University of Bergen

Department of Mathematics

Hidden Markov and Hidden Semi-Markov models on Financial Timeseries

Author:
Emil Lund Eilertsen

Supervisor:
Antonello Maruotti

Master's Thesis in Statistics

Financial Theory and Insurance Mathematics

June 16th 2020

Abstract

Applications related to Financial Econometrics like risk measurement and many other financial indicators rely on a suitable modeling of the distributional and temporal properties of the daily return series of stocks. Financial data rarely comes from a homogeneous population and most often there is an underlying latent (hidden) structure that affect the observable variables. Hidden Markov models have been widely applied in financial fields due to the features in describing these underlying structures of the financial data allowing to measure components distribution with several underlying components capturing the underlying regimes in the data. The Hidden Markov model are often used to model daily returns and to infer the hidden state of financial markets, and has been shown to reproduce most of the stylized facts about stock return series. A notable exception is the inability of the HMM's to reproduce one ubiquitous feature of such time series, namely the slow decay in the autocorrelation function of the absolute and squared returns. It is shown that this stylized fact can be described much better by means of hidden semi-Markov models.

In this thesis we present hidden semi-Markov models (HSMM) with different combinations of sojourn time distribution (SD) and emission distribution (ED) in order to improve univariate risk measures such as Value at Risk (VaR) and Expected Shortfall (ES). We use the widely used model selecting criteria AIC and BIC in order to find the best fitted models on each of the three dataset used, SP500, ESTX50 and FTSE. We further use these models to see how well they reproduce the stylized facts of daily stock return series, and then we examine how well the models reproduce the original data comparing the empirical cumulative distribution function (ECDF) of the original data with the ECDF of the fitted models.

Keywords: Latent variable models; Hidden Markov model; Hidden semi-Markov model; EM-algorithm; Model selection; Daily return series; Stylized facts; Risk measurements;

Acknowledgement

I would like to thank my supervisor Antonello Maruotti for his great guidance and great supervision during the writing of this thesis. Furthermore, I would like to express my gratitude to my fellow students for many interesting discussions and friendship throughout the study period, the faculty members and staff at the Department of Mathematics, especially Kristine Lysnes, the senior consultant, for always being helpful if there ever was some uncertainties of any kind

Finally, I would like to thank family and friends for their moral support and words of encouragement throughout the study period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sometimes traditional statistic models fail to capture the essence of given data. The reason may be because the observations that are being analysed does not come from a homogeneous population, or maybe there is some underlying latent structure that affect the observable variables. This means that it's unlikely that all of the observations in our sample have the same set of parameter values. A Latent (hidden) variable is a variable that is not directly observed but rather inferred through a mathematical model from other observable variables. When you have a mathematical model whose goal is to explain the observable variables by using the latent variables, the model is called a latent variable model. Latent variable models include a large specter of different models that can be applied to different fields in statistics. Three examples of latent variable models that I will talk about in this thesis is Compound distribution, Finite mixture models and hidden Markov models. Compound distribution can be seen as a continuous mixture model, finite mixture model can be seen as a non-parametric approach for the mixing distribution used in the compound approach and the Hidden Markov model can be seen as an extension of finite mixture model to time-dependent data analysis. Hidden Markov models are models in which the distribution that generates an observation depends on the state of an underlying and unobserved Markov process. They provide flexible general-purpose models for univariate and multivariate time series, especially for discrete valued series, including categorical series and timeseries. The main model I will cover in this thesis is the HSMM, which is an extension of the HMM. The underlying process of the HSMM is an semi-Markov chain that allows one to utilize more flexible sojourn time distributions than that of a HMM model where the state duration implicitly is a geometric distribution. Contrary to the HMM model, for each state in the HSMM model there is a variable duration d modelled by the survivor function (eq 3.2), which is the key to extend the original algorithms that say there is a change of state immediately after the last observation.

The growing popularity of HMM's in the past decades has led to numerous papers on applications to real-world problems, and also an increased interest in computational aspects. In order to estimate the parameters of the HMM model, one must employ maximum likelihood (ML) parameter estimation, mostly by implementing a expectation maximization algorithm or alternatively a numerical maximization algorithm. In the case of the HSMM's, the situation is, however, slightly different. The main difference, compared to HMM, is that they allow for a greater flexibility for the choice of the sojourn time distributions. Unfortunately, this flexibility comes with a much higher computational burden. In order to make this class of models accessible to a larger number of researchers, there exists some packages, including the package hsmm introduced by Bulla et al. [2010] and the package mhsmm introduced by O'Connell and Højsgaard [2011], which is the one used in this thesis. The mhsmm package implements all the most important algorithms required for working with HSMM and HMM, which includes estimation of parameters and prediction.

Applications related to Financial Econometrics like risk measurement and many other financial indicators rely on a suitable modeling of the distributional and temporal properties of the daily return series of stocks, indices or other assets. The normal distribution with stationary parameters has often been chosen to model daily return series in financial theory. However, the lack of the Normal distributions ability to capture skewness and kurtosis, which is well known to be present in financial timeseries, makes it not an appropriate model for financial timeseries. After a paper by Fama [1965], which observed more kurtosis and higher peaks contradicting the assumption of normality, many authors proposed solutions to overcome this drawback. Blattberg and Gonedes [1974] preferred the t distribution and many other authors has proposed different distributions, like Eling [2012] who preferred skew-normal and skew-t distribution.

Many different stylized facts have been established for financial returns, see e.g. Granger and Ding [1995a] and Granger and Ding [1995b]. Rydén et al. [1998] showed the ability of a hidden Markov model (HMM) to reproduce most of the stylized facts of daily return series introduced by Granger and Ding [1995a] and Granger

and Ding [1995b]. In an HMM, the distribution that generates an observation depends on the state of an unobserved Markov chain. Rydén et al. [1998] found that the one stylized fact that could not be reproduced by an HMM was the slow decay of the autocorrelation function (ACF) of squared and absolute daily returns, which is of great importance in financial risk management. Rydén et al. [1998] considered this stylized fact to be the most difficult to reproduce with an HMM. According to Bulla and Bulla [2006], the lack of flexibility of an HMM to model this long-memory property can be explained by the geometrically distributed sojourn times in the hidden states. This led them to consider The HSMM model in which the sojourn time distribution is modeled explicitly for each hidden state. Bulla and Bulla [2006] found that an HSMM with negative-binomially distributed sojourn times was better than the HMM at reproducing the long memory property of squared and absoloute daily returns.

In this thesis, however, we will mainly focus on the Hidden semi-Markov models, and try different combinations of sojourn distribution (SD) and emission distribution (ED) in order to find the best fitted models on the datasets SP500, ESTX50 and FTSE based on the model selection criterias AIC and BIC. We will compare these different combinations of models to hidden Markov models with Normal ED and number of states varying from K = 2 to 6. The sojourn distributions available in the mhsmm package is shifted Poisson SD and Gamma SD. The emission distributions available in the mhsmm package is the Normal ED and the Poisson ED, and a user-defined extension for Multivariate Normal distribution presented in O'Connell and Højsgaard [2011]. In this thesis, we want to investigate whether the t-distribution, the skew-t distribution and skew-normal distribution as emission distributions will improve the fit of the three different datasets. In order to do that we must extend the mhsmm package and write our own user-defined emission distribution for each of the distributions mentioned above. This is more thoroughly explained in chapter 7. Furthermore, we will combine the two sojourn distributions together with the four introduced emission distributions. The number of states in each fitted model vary from K = 2 to 6 which results in 40 different HSMM models that will be tested on each of the three datasets. In addition we will investigate HMM models with Normal ED and the default SD, namely the geometric distribution, on K = 2 to 6 states, which results in 5 different models. Based on the values obtained from the model selection criterias, we will continue the analysis chapter with the three best fitted models on each dataset. The best HMM model, the best HSMM model with Gamma SD and the best HSMM model with shifted Poisson SD for each dataset. We will use the best fitted models to improve the estimates of univariate risk measures, Value at Risk (VaR) and Expected Shortfall (ES), which is explained in chapter (8.3.1). We will further check for the stylized facts of stock return series, and show that the stylized fact can be described better by means of HSMM, chapter (8.3.3). Then we will do an in-sample analysis to see how good the different models perform in reproducing the original data, which is explained in chapter 8.3.4. As an additional analysis, we will do an component distribution analysis of the HSMM with Gamma SD and Normal ED for K = 3 states on the ESTX50 dataset to show how one can interpret and identify different periods of volatilities, explained in chapter (8.3.2).

Summarized, this thesis is divided into three parts. Following this introductory part is Part Two, which presents background information necessary for the empirical analysis. The organization in Part Two is structured as follows: Chapter 2 gives in introduction to latent variable models where three latent variable models mentioned above are introduced. Chapter 3 gives a thorough explanation of the HSMM model and its algorithms. Chapter 4 introduce the four different distributions used as the emission distribution in the HSMM framework. Chapter 5 gives a brief introduction to Financial Returns. Chapter 6 describes the importance of Risk Management and univariate risk measures and chapter 7 describe the extension of the mhsmm package where the four different distribution introduced in chapter 4 are implemented as emission distributions. Part Three provides the empirical results. This part is divided into three sub-chapters, beginning with Chapter 8.1, which covers the descriptive statistics related to the dataset used in this thesis. Chapter 8.2 which describes model selection criteria and Chapter 8.3, which give us an insight into the Empirical Analysis and results. The analysis presented in chapter 8 is performed in R, a programming language for statistical computing and graphics. Chapter 9 is a summary of this thesis, including the conclusions we have drawn from our research and suggests several ideas for related future work. Following these concluding chapters are several appendices and at the end is the bibliography.

Chapter 2

Latent Variable models

A Latent (hidden) variable is a variable that is not directly observed but rather inferred through a mathematical model from other observable variables. When you have a mathematical model whose goal is to explain the observable variables by using the latent variables, the model is called a latent variable model. Latent variable models can be used to answer questions about the latent constructs, regarding for example how their means and standard deviations vary between populations, and also how different constructs are associated with each other. Any model that relates some kind of latent structure to an observed structure could be called a latent variable model, and the possibilities regarding the dimensionality and form of these structures are endless. Unlike, for example, the normal family of distributions where the mean and standard deviation are known, the distribution is normal and the probability of any future observation lying in a given range is known. This is not the case for latent variable Models. In this chapter I will introduce three latent variable models; Compound distribution, Finite mixture and hidden Markov models. I will first introduce the compound distribution, which can be seen as continuous mixture models. Then I will introduce the Finite mixture model, which can be seen as a non-parametric approach for the mixing distribution used in the compound approach. And last, I will introduce the Hidden Markov model, which can be seen as an extension of finite mixture model to time-dependent data analysis.

2.1 Compound Distribution

Definition: wikipedia
A compound probability distribution is the probability distribution that results from assuming that a random variable X is distributed according to some parameterized distribution F with an unknown parameter theta that is again distributed according to some other distribution G. The resulting distribution H is said to be the distribution that results from compounding F with G

In this chapter, the theory is based on the following articles books: Cai and Garrido [1999], Ma [2010], Lin [2006], Pitts [1994], Willmot and Lin [2001] and Shevchenko [2010] which gives a thorough introduction and explanation on the compound distribution.

The most important job in actuarial risk modelling is accurately fitting the tails of the insurance losses. In particular, the losses in the right tail, though rare in frequency, are indeed the ones that have the most impact on the operations of an insurer as it could lead to possible bankruptcy of the company. In such circumstances, heavy tailed distributions such as Pareto, lognormal, Weibull and gamma distribution have been shown to be reasonable competitive. However, each of these distributions covers different behaviour of losses. While Pareto does not provide a reasonable fit when the density of the data is hump shaped, lognormal, Weibull, and gamma distributions better cover the behaviour of small losses but fail to cover the behaviour of large losses.

A desired model is a model that account for all the peculiarities of the loss data discussed above. Compound distribution is an approach that combines different distributions to obtain a new probability distribution that gives a more precise and accurate analysis. Continuous compound distribution is a desirable choice if you want to improve the tail behaviour of any unimodal distribution with positive support.

In continuous compound distribution the variability-related parameter is scaled by a suitable arbitrary variable. In particular, a 2- parameter unimodal hump-shaped model is considered and is defined on a positive support, that is, values on the positive real line. The model is parameterized with respect to 2 parameters, $\gamma > 0$ and $\theta > 0$. The γ parameter is closely related to the variability in the distribution and θ is the mode.

The γ parameter is then scaled by some parameterized mixing distribution that takes on values on the whole or only parts of the positive real line and is also dependent on a single parameter that governs the behaviour of the tail. The result is a 3-parameter compound distribution which gives more flexibility to the tails of this new conditional distribution. Various classes of models can be considered for the mixing distribution to give the best result and the resulting model guarantees unimodality in the parameter theta and smoothness

Compound distributions are widely used in modelling the aggregate claims in an insurance portfolio. Compounding of probability distribution is a method to obtain new probability distributions by combining the primary distribution, which is the distribution of a counting variable N (number of claims), and the secondary distribution, which is the distribution of an individual claim amount X_i.

The compounding of probability distributions enables us to obtain both discrete as well as continuous distribution If we for instance consider the random variable S, where S is of the form $S = X_1 + X_2 + ... + X_N$, the random variable S is said to have a compound distribution where the number of terms N is uncertain, the random variables X_i are independent and identically distributed and each X_i is independent of N. The random sum S represents the aggregate claims amount, the primary distribution (number of claims, N) represents the claim frequency distribution and the secondary distribution (claim amounts, X_i) represent the claim severity distribution.

Let $\{X_1, X_2, \ldots, X_N\}$ be an independent and identically distributed sequence of positive random variables, independent of N, with common distribution function $F(x) = P\{Xi \leq x\}$, $x \geq 0$, where X_i is an arbitrary variable from the sequence $\{X_1, X_2, \ldots, X_N\}$. $\{p_n; n = 0, 1, 2, ..., N\}$ specifies the number of claims distribution and $F(x), x \geq 0$ is the individual claim amount distribution. Also, let $F^{*n}(x) = P\{\sum_{i=1}^{N} X_i \leq x\}$ for n = 1,2,...,N be the n-fold convolution of $F(x)$ and $\overline{F}^{*n}(x) = 1 - F^{*n}(x) = P\{\sum_{i=1}^{N} X_i > x\}$ be the tail. The distribution of the random sum $S = \sum_{i=1}^{N} X_i, H(x) = P\{S \leq x\}$ with the convention that S = 0, if N = 0, is called a compound distribution and is given by

$$H_S(x) = \sum_{i=1}^{\infty} p_n F^{*n}(x), x \geq 0 \tag{2.1}$$

where $F^{*0}(x) = 1$, and therefore

$$\overline{H_S}(x) = \sum_{i=1}^{\infty} p_n \overline{F}^{*n}(x), x \geq 0. \tag{2.2}$$

In general, evaluation of the tail of the aggregate claims $\overline{H_S}(x)$ is difficult due to the presence of the convolutions. But if the individual claim amount distribution F(x) is closed under convolution, as in compound Binomial distribution, simplification occurs. It is shown that the formulation of the compound Binomial distribution provides closed forms for the marginal probabilities if the Laplace transform of the mixing distribution may be written in closed form.

The construction of formula (2.1) is unfortunately so intricate that the direct computation of $H_S(x)$ is tractable only in special cases, as mentioned above. An approximation is therefore necessary in order to evaluate a compound distribution. Several methods such as moment-based analytic approximations has been proposed for this problem. These analytic approximations perform relatively well for compound distributions with small skewness. However, when the skewness increase, these approximation methods is no longer a good alternative.

Another approach is numerical evaluation procedures. These methods are based on simulating the frequency and severity distributions and is often necessary for most compound distributions in order to obtain a required degree of accuracy. However, a simulation algorithm is often ineffective and requires a great capacity of computing power.

The first three moments of the compound distribution are as follows:
The expected aggregate claims $E[S]$ is:

$$\begin{aligned} E[S] &= E_N[E(S \mid N)] \\ &= E_N[E(X_1 + ... + X_N \mid N)] \\ &= E_N[NE(X)] \\ &= E[N]E[X] \end{aligned} \tag{2.3}$$

The expected value of the aggregate claims is the product of the expected number of claims and the expected individual claim amount.

The variance of the aggregate claims Var[S] is:

$$\begin{aligned} Var[S] &= E_N[Var(S \mid N)] + Var_N[E(S \mid N)] \\ &= E_N[Var(X_1 + ... + X_N \mid N)] + Var[E(X_1 + ... + X_N \mid N)] \\ &= E_N[NVar(X)] + Var[NE(X)] \\ &= E[N]Var[X] + Var[N]E[X]^2 \end{aligned} \tag{2.4}$$

The skewness of any random variable Z is defined as:

$$\gamma_Z = E[(\tfrac{(Z-\mu_Z)}{\sigma_Z})^3] = \sigma_Z^{-3} E[(Z-\mu_Z)^3]$$

Since $\Psi_Z^{(3)}(0) = E[(Z-\mu_Z)^3]$, we have $\gamma_Z = \sigma_Z^{-3}\Psi_Z^{(3)}(0)$ and $\Psi_Z^{(3)}(0) = \sigma_Z^3\gamma_Z$ We have that $\Psi_Y(t) = \Psi_N[\Psi_X(t)]$. So by taking the third derivative of of $\Psi_Y(t)$ and evaluating at t = 0 we get:

$$\Psi_Y^{(3)}(0) = \gamma_N\sigma_N^3\mu_X^3 + 3\sigma_N^2\mu_X\sigma_X^2 + \mu_N\gamma_X\sigma_X^3$$

Thus, the following is the skewness of the aggregate claims Y:

$$\gamma_Y = \frac{\gamma_N\sigma_N^3\mu_X^3 + 3\sigma_N^2\mu_X\sigma_X^2 + \mu_N\gamma_X\sigma_X^3}{(\mu_N\sigma_X^2 + \sigma_N^2\mu_X^2)^{\frac{3}{2}}} \tag{2.5}$$

2.2 Finite Mixture

In finance, the form of the distribution of stock returns is a crucial assumption for modelling and analysing the dataset in a best way possible. The distribution of stock returns tends to have both significant kurtosis and significant skewness. In these situations, the Normal distribution, though widely used in the financial world, lack the ability to sufficiently model the data. There are, however, several distributions that allow to regulate for both kurtosis and/or skewness but fail to capture the normality aspect of the stock returns. Finite mixture models have been widely applied in financial fields due to the features in describing the complex system on the financial data analysis, allowing to measure a two or more component distribution and in estimating a mixing probability in the data. The advantage of finite mixture models include that they maintain the tractability of normal distribution while having finite higher order moments and can capture the excess kurtosis and skewness which is an important aspect of analysing financial data. Finite mixture models are also useful in measuring heavy-tailed densities, examine heterogeneity in a cluster analysis, analysing a mixture of univariate distribution and estimating a mixing probability in the data, which also are important aspects in modelling financial stock returns.

Financial stock returns can be divided into several periods, that is, a bull market, a bear market and a sidewalk market. The bull market can be defined as having a higher frequency of positive returns, the bear market can be defined as having a lower frequency of positive returns and the sidewalk market can be defined as having more or less the same frequency of both positive and negative returns. Combining these markets, a finite mixture model with K different components corresponding to each market, each distributed as e.g. a Normal distribution with unknown mean and variance is a reasonable model to capture the behaviour of stock returns.

In this chapter, the theory is based on the following articles/books: McLachlan et al. [2018], McLachlan and Peel [2000], Picard [2007], Phoong and Ismail [2014], Chung et al. [2004], Marin et al. [2005] and Kon [1984] which gives a thorough introduction and explanation on the finite mixture model

Most of statistical models assume that a sample of observations comes from the same distribution. Sometimes, however, it may not be true, since the sample may be drawn from numbers of distinct populations in which the populations are not identified. In this situation the homogeneity assumption is violated, and the Finite Mixture (FM) model is a good method to handle these kinds of situations.

A Finite Mixture (FM) model is a useful tool in modelling heterogeneous data with a finite number of unobserved sub-populations. In finite mixture models the total set of data that we have observed is a mixture of underlying subgroups, and within each subgroups we have a distribution of a particular type. A common type of finite mixture models is normal mixture model, were we assume that there is a normal distribution within each subgroup. Because it does involve a formal statistical model, the finite mixture-modelling framework can incorporate lots of models that we're already familiar with. Finite mixture models can be applied to classification, clustering, and pattern identification problems for independent data, and could also be used for longitudinal data to describe differences in trajectory among these subgroups. However, due to the computational convenience, the most types of FM models are based on the normality assumption, which may be violated in certain real situations.

When the data we are interested in contains a tail that is longer and/or heavier than the tail of the normal distribution, as well as have atypical observations, the t-distribution is considered a good alternative. It provides a more robust approach of fitting mixtures, as the observations that are atypical of a component are given reduced weight in the calculation of its parameters, and computes fewer extreme estimates of the posterior probabilities of the component membership of the mixture model. When the data involve asymmetric features, the use of symmetric distributions such as normal distribution and t-distribution can be very misleading when handling the skewness in the data. Asymmetric distribution-based mixture models like the Skew-Normal (SN) and Skew-t (ST) are much better suited when modelling data with asymmetry, heavy tails, and the presence of outliers. By adding additional shape/skewness parameters, the SN distribution can provide a more appropriate density estimation to fit the asymmetric observations, compared to the normal mixtures. The ST distribution has advantages in modelling data with both asymmetry and heavy tails simultaneously. Compared to the SN distribution, the ST distribution has extra parameters, degrees of freedom and shape/skewness parameter.

The probability density function, or probability mass function in the discrete case of a finite mixture distribution of a p-dimensional random vector y, takes the form:

$$f(y) = \sum_{i=1}^{K} \pi_i f_i(y) \tag{2.6}$$

where the mixing proportions π_i are non-negative and $\sum_{i=1}^{k} \pi_i = 1$, and where the f_i's are the component densities.When we specify a parametric form $f_i(y_j; \theta_i)$ for each component density, we can fit the parametric mixture model

$$f(y_j; \Psi) = \sum_{i=1}^{K} \pi_i f_i(y_j; \theta_i) \tag{2.7}$$

by maximum likelihood with the EM-algorithm. Here $\Psi = \{\pi_1, \pi_2, ..., \pi_K, \zeta\}$ and $\zeta = \{\theta_1, \theta_2, ..., \theta_K\}$ is all the parameters of the mixture model where θ_i is known a priori to be distinct and $f_i(y_j; \theta_i)$ is the j'th component density for observation y_i with parameter vector θ_i. Assuming that K is fixed, the model parameters Ψ are usually unknown and must be estimated. The likelihood function corresponding to equation (2.7) is given by

$$L(y_1, y_2, ..., y_n, \Psi) = \prod_{j=1}^{n} \left(\sum_{i=1}^{K} \pi_i f_i(y_i; \Psi) \right) = \sum_{i=1}^{K} \left(\prod_{j=1}^{n} \pi_j \prod_{j=1}^{n} f_j(y_j; \Psi_j) \right) \tag{2.8}$$

And the log likelihood function corresponding to (2.7) is given by

$$logL(\Psi) = log \sum_{i=1}^{K} \left(\prod_{j=1}^{n} \pi_j \prod_{j=1}^{n} f_j(y_j; \Psi_j) \right) = \sum_{i=1}^{K} \sum_{j=1}^{n} (log\pi_j + log f_j(y_j, \Psi_j)) \tag{2.9}$$

Direct maximization of this log-likelihood equation encounter some difficult computations, so the maximum likelihood estimator is obtained by using the EM-algorithm. The maximum likelihood estimate of Ψ is obtained by solving this equation:

$$\frac{\partial logL(\Psi)}{\partial \Psi} \tag{2.10}$$

where $L(\Psi)$ denotes the likelihood function for the mixture model. Solution of equation (2.10) can be found by using the Em-algorithm. Once $\hat{\Psi}$ is obtained, estimates of the posterior probabilities of the population membership can be formed for each observation to give a probabilistic classification of the data.

When the components densities are not fixed they have to be inferred from the available data along with the mixing proportions and the parameters in the specified forms for the component densities. One way of generating a random vector Y_j with a K-component mixture density $f(y_j)$ given by:

$$f(y) = \sum_{i=1}^{k} \pi_i f_i(y),$$

is as follows. Let Z_j be a categorical random variable taking the values $1, ..., K$ with probabilities $\pi_1, ..., \pi_K$, respectively and suppose that the conditional density of Y_j given $Z_j = i$ is $f_i(y_j \mid i = 1, ..., K)$. Then the unconditional density of Y_j (the marginal density) is given by $f(y_j)$. The variable Z_j can be thought of as the component label of the feature vector Y_j.

A mixture model is able to model quite complex distributions through an appropriate choice of its components to represent accurately to local area of support of the true distribution. It can thus handle situations where a single parametric family is unable to provide a satisfactory model for the local variations in the unobserved data. One of the advantages of FM models is that both maximum likelihood method and Bayesian approach can be applied to not only estimate model parameters, but also evaluate probabilities of subgroup membership simultaneously .

The optimal number of mixture components selection is an important but difficult problem in FM models. One approach to determine this is to compare the information criteria, such as Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC) and Sample-Size Adjusted BIC. However, most of these criteria are very sensitive to sample size, and favour highly parameterized models. Thus, it is suggested that these information criteria should be considered with other evidence. For optimal results it is necessary to apply different criteria simultaneously to determine the optimal number of components for FM models.

2.2.1 Computational aspects and inference

There is a large amount of estimation methods that are being used to estimate the parameters of a finite mixture model. That include methods such as EM- algorithm with classification EM and stochastic EM, direct numerical maximization, hybrid approaches, Bayesian methods etc. The question that arises is which estimation method that should be chosen for the estimation of parameters of a mixture distribution. The most common one is the EM-algorithm, however, there are several drawbacks in using likelihood approach. The EM algorithm leads to a local maximum, so in order to find a global maximum, a grid of many different starting points is needed. The sample size also needs to be very large because the maximum likelihood method is based on the asymptotic theory.

Another approach that also has gotten a lot of attention is the Bayesian approach. Bayesian sampling approach is an approach which provides a richer inference than the ML approach in that it can address the issue of parameter uncertainty through full posterior distribution. Bayesian approach simulates random draws of parameters from a posterior distribution using Markov chain Monte Carlo (MCMC). MCMC may produce estimates and credible regions for the parameters without appealing to large sample approximations. Bayesian method is the only sensible method to use if the number of mixture components is allowed to vary. From a computational standpoint, simulating draws from the posterior distribution by MCMC is no more difficult than ML estimation by EM. With MCMC, however, many new issues arise. Maybe the most troubling aspect of MCMC in a finite mixture framework is that the component labels may switch during an MCMC run.

In this section I will show how parameters is estimated using the EM algorithm. The EM algorithm has three essential requirements, that is, the choice of reasonable initial values, an iterative algorithm that defines the new estimates and an appropriate stopping criterion. The EM algorithm consists of two steps, The expectation step (E-step) and the maximization step (M-step). The basic idea is to associate an incomplete dataset to a complete dataset, which makes the ML estimation more tractable.

To compute equation 2.10, an unobservable or missing data vector $z = (z_1, ..., z_n)$ is introduced, where each $z_{ij} = (z_j)_i$ is a k-dimensional vector of indicator variables that are zero or one according to whether the i'th observation did or did not arise from the j'th component of the mixture.

So the complete data log likelihood is

$$logL_c(\Psi) = \sum_{i=1}^{K} \sum_{j=1}^{n} z_{ij}(log\pi_j + logf_j(y_j, \Psi_j)) \tag{2.11}$$

The E-step requires averaging $logL_c(\Psi)$ over the conditional distribution of z given the observed data vector y, using the current fit for the vector of unknown parameters Ψ. As $logL_c(\Psi)$ is linear in the unobservable data z_{ij}, the E-step simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data y, where Z_{ij} is the random variable corresponding to z_{ij}. This yields the Q-function given by

$$Q(\Psi;\Psi^k) = \sum_{i=1}^{k}\sum_{j=1}^{n} \tau_i(\gamma_j;\Psi^k)(log\pi_j + logf_j(y_j,\Psi_j)) \tag{2.12}$$

where

$$\begin{aligned} \tau_i(\gamma_j;\Psi^{(k)}) &= E_{\Psi^{(k)}}(Z_{ij} \mid y) \\ &= P_{\Psi^{(k)}}(Z_{ij}=1 \mid y_j) \\ &= {\pi_i}^{(k)} f_i(y_j;{\theta_i}^{(k)})/f(y_j;\Psi^{(k)}) \end{aligned} \tag{2.13}$$

is the posterior probability that the j'th observation y_j belongs to the i'th component. $E_{\Psi^{(k)}}$ denote the expectation and $P_{\Psi^{(k)}}$ denote the probability.
The M-step requires the global maximization of $Q(\Psi;\Psi^k)$ with respect to Ψ over the paramater space to give the updated estimate Ψ^{k+1}. The updated estimate of the i'th mixing proportion π_i is then given by

$${\pi_i}^{k+1} = \sum_{j=1}^{n} \tau_i(\gamma_j;\Psi^k). \tag{2.14}$$

The updated estimate of the vector ζ containing the distinct parameters in the component densities satisfies the equation

$$\sum_{i=1}^{k}\sum_{j=1}^{n} \tau_{ij}(\gamma_j;\Psi^k)\frac{\partial logf_i(\gamma_j,\theta_i)}{\partial\zeta} = 0 \tag{2.15}$$

The E-step and M-step are alternated repeatedly until the difference

$$L\left(\Psi^{k+1}\right) - L\left(\Psi^k\right) \tag{2.16}$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $L\left(\Psi^k\right)$

2.3 Hidden Markov Model

A consistent challenge of financial traders is the frequent behaviour of financial markets. One period can experience a lot a negative return frequencies, another period can experience a lot a positive return frequencies while other periods might be relatively stable. These various periods, known as market regimes, lead to adjustments of asset returns via shifts in their means, variances/volatilities and autocorrelation, which impact the effectiveness of time series methods that rely on stationarity. In particular, it can lead to dynamically varying correlation, excess kurtosis, heteroskedasticity as well as skewed returns. Hidden Markov models have been widely applied in financial fields due to the features in describing these complex systems of the financial data analysis and allowing to measure components distribution with several underlying components capturing the underlying regimes in the data. In other words, hidden Markov models are well suited to this task as they involve inference on "hidden" generative processes via "noisy" indirect observations correlated to these processes. In this instance the hidden, or latent process is the underlying regime state, while the asset returns are the indirect noisy observations that are influenced by these states. The theory of this chapter is based on the following articles/books: Zucchini et al. [2016], Tolver [2016], Jurafsky and Martin [2019], Fosler-Lussier [1998], Rabiner [1989], Stamp [2018] and Nguyen [2018] which gives a thorough introduction and explanation on the Hidden Markov model. The texts also provides for further reading.

Before I go into details on the Hidden Markov model, let me first introduce Markov chains, the simplest Markov model. First of all, a Markov chain is a stochastic process that models the state of a system with a random variable that changes through time.

A Markov chain is a discrete-time process were the future behaviour, given the present and the past, only depends on the present behaviour. All the states before the present state has no impact in predicting the future behaviour. For example, if we want to predict tomorrow's weather using a Markov chain, you could only look at today's weather and not at the weather in the days before. By definition, the probability that the stochastic process X of being in a state j depends only on the previous state, and not on any other states that occurred before that.

$$P(X_t = j \mid X_1 = i_1, ..., X_{t-1} = i_{t-1}) = P(X_t = j \mid X_{t-1} = i_{t-1}) \tag{2.17}$$

This is the probability that the process X is in state j at time t, given that the previous state was i_{t-1} at time $t-1$. All the other states that happened longer back in time are of no interest

Consider a Markov chain at the discrete time points {0, 1, 2,...}. The Markov chain is characterized by the three following components.

- A set of states, $S = \{S_1, S_2, ..., S_N\}$
- The transition probabilities, p_{ij}, between each state. P_{ij} is the probability that the Markov chain is at the next time point in state j, given that it is at the present time point at state i. The process can remain in the current state, and this occurs with probability p_{ii}
- An initial probability distribution over states. π_i is the probability that the Markov chain will start in state i

The matrix P with elements p_{ij} is called the transition probability matrix of the Markov chain. Note that the definition of the p_{ij} implies that the row sums of P are equal to 1. This is because of that the total probability must equal to 1. A Markov chain is useful tool when we need to compute a probability for a sequence of observable events. In many cases, however, the events we are interested in are hidden which means we don't observe them directly.

That's where the Hidden Markov model comes into the picture. A Hidden Markov model (HMM) is a stochastic process where we have an underlying, invisible Markov chain where each state of the Markov chain generates only one out of K possible observations. These observable output observations are state dependent and visible to us. So we can say that a HMM is a Markov process that is split into two components, an observable and an unobservable (hidden) component. The process S_t which represents the underlying unobserved process of the HMM fulfils, just as the Markov chain, the Markov property;

$$P(S_t = j \mid S_1 = i_1, ..., S_{t-1} = i_{t-1}) = P(S_t = j \mid S_{t-1} = i_{t-1}), \tag{2.18}$$

meaning that the probability of the process S of being in a state j depends only on the previous state i_{t-1}.

Let $\pi_k = P(S_1 = k)$ be the initial probability of state k, $k = 1, ..., K$. Let

$$P_{jk} = P(S_t = k \mid S_{t-1} = j) \tag{2.19}$$

denote the transition probability, that is, the probability of being in state k at time t given that previous state was j at time $t-1$, We must also have that $\sum_{k=1}^{K} P_{jk} = 1$ and $P_{jk} > 0$. The initial probabilities $\pi_k = (\pi_1, \pi_2, ..., \pi_k)$ together with the transition probability matrix P, where P_{ij} is the elements of the matrix, govern the state switching behaviour of the chain. The number of time spent in each state before jumping to the next state is called the sojourn time. The probability of spending u consecutive time steps in state i under this model is

$$\begin{aligned} d_i(u) &= P(S_{t+u+1} \neq i, S_{t+u} = i, S_{t+u-1} = i, ..., S_{t+2} = i \mid S_{t+1} = i, S_t \neq i) \\ &= P_{ii}^{u-i}(1 - P_i) \end{aligned} \tag{2.20}$$

We call $d_i(u)$ the sojourn density. Hence the sojourn time is geometrically distributed for any Markov chain, and the most likely sojourn time for any state is equal to 1. One weakness of the HMM is the lack of adaptability to different sojourn time distributions, since it is based on a hidden Markov chain whose sojourn times follow a geometric distribution. This is not always desirable and limits the range of possible applications

One example concerning the HMM is that you can think of X_t as the market price of stock and S_t as an unobserved economic factor process that influences the fluctuations of the stock price. We are ultimately interested in modelling the observed stock price fluctuations, not the fluctuations in the unobservable factor process. But by including the unobserved process in the calculations we might be able to build a model that

more precise capture the statistical properties of the observed stock prices. It should be noted that even though St is a Markov process, typically the observed component X_t would not be a Markov process itself. Hidden Markov models can thus be used to model non-Markov behaviour (for example the stock price), while retaining many of the mathematical and computational advantages of the Markov setting

The probability of an output observation, X_t, depends only on the state that produced the observation, S_t, and not on any other states or any other observations. Let $S = (S_t, t = 0, ..., T)$ denote the sequence of unobserved random variables, each with a finite state space {1,..., J}, and let $X = (X_t, t = 1, ..., T)$ denote a corresponding set of observed random vectors. The process $\{X_t\}$ represents the state-dependent process of the HMM and fulfils the conditional (on the hidden states) independence property

$$\begin{aligned} &P(X_t = x_t \mid X_1 = x_1, ..., X_{t-1} = x_{t-1}, S_1 = s_1, ..., S_t = s_t) \\ &= P(X_t = x_t \mid S_t = s_t) \end{aligned} \tag{2.21}$$

This is called the emission probability. The probability that you will see x_t given that at the same time you are in state s_t.

The hidden Markov model has the functional form

$$P(X, S) = P(X \mid S)P(S) \prod_{t=1}^{T} P(X_t \mid S_t) \prod_{t=1}^{T} P(S_t \mid S_{t-t}) \tag{2.22}$$

Given a specific observation sequence we want to calculate, we must first compute the joint probability of being in a particular hidden state sequence S_t and generating a particular sequence X_t of observable events. Then we must compute the total probability of the observations just by summing over all possible hidden state sequences.

$$P(X) = \sum_{S} P(X, S) = \sum_{S} P(X \mid S)P(S) \tag{2.23}$$

For a Hidden Markov model with an observation sequence of T observations and N hidden states, there are N^T possible hidden sequences. In real life situations, even if the length of the sequences N and T are moderate, N^T becomes a very large number. This makes it hard for us to compute the total observation likelihood.

Fortunately there exist a couple of algorithms we can use that makes it easier for us to compute the probability. The Forward Algorithm, The Backward Algorithm and The Viterbi Algorithm are three of the most used algorithms to compute the probability of a given observation sequence.

2.3.1 Computational aspects and inference

In many applications of HMM it is difficult to know how to design the transition and observation kernels and the initial measure to obtain the best result. This is especially true in modelling financial time series models, where the design of a hidden Markov model should explain in a best way possible the observation process so the latter possess the desired statistical properties.

It is therefore essential to develop statistical inference techniques which allow us to design and calibrate our hidden Markov model to match observed real-world data. It should be noted that in this setting we may not have much, if any, a priori knowledge of the structure of the unobserved process.

For a HMM to be useful in real world application, the following three problems must be solved.

The evaluation problem: *we observe a finite number of observations $x_0, ..., x_N$, and we wish to find the probability that the observations is generated by the hidden model.*

The Solution to this problem is solved by using the forward algorithm. The forward algorithm is an algorithm that stores intermediate values in a table as it builds up the probability of the observation sequence. By summing over the probabilities of all hidden state paths that generates the specific observation sequence, the forward algorithm computes the observation probability and creates single forward trellis from each of the paths.

Each cell of the forward algorithm trellis $\alpha_t(j)$ represents the probability of being in state j after seeing the first t observations, given the model $\lambda = (A, B, \pi)$, where A is the transition probabilities, B is the emission probabilities and π is the initial probabilities. The value of each cell $\alpha_t(j)$ is computed by summing over the probabilities of every path that could lead us to this cell.

Formally, each cell expresses the following probability:

$$\alpha_t(j) = P(x_1, x_2, ..., x_t, s_t = j \mid \lambda) \tag{2.24}$$

where $s_t = j$ is the t'th state you are at. We compute this probability $\alpha_t(j)$ by summing over the extensions of all the paths that lead to the current cell. For a given state s_j at time t, the value $\alpha_t(j)$ is computed as

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(x_t) \tag{2.25}$$

Equation (2.24) extends the previous paths to compute the forward algorithm and consists of three factors. That is, $\alpha_{t-1}(i)$ the previous forward path probability from the previous time step, a_{ij} the transition probability from previous state s_i to current state s_j and $b_j(x_t)$ the state observation likelihood of the observation symbol x_t given the current state j.

The same result can be obtained by using the Backward algorithm, given by

$$\beta_t(i) = P(x_{t+1}, x_{t+2}, ..., x_T, s_t = i \mid \lambda) \tag{2.26}$$

In contrast to the forward probability, the backward probability is the probability of seeing the observations from time (t+1) to the end (T), given that we are in state i at time t. And as in the case of $\alpha_t(j)$ there is a recursive relationship which can be used to calculate $\beta_t(i)$ efficiently

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(x_{t+1}) \tag{2.27}$$

The decoding problem: *we have sequence of observation $x_0, ..., x_N$, and wish to find what the most likely state sequence in the model that produced the observations.*

The solution to this problem is solved by using the Viterbi algorithm. The Viterbi algorithm is almost identical to the forward algorithm, the only difference is the Viterbi algorithm takes the max over the previous path probabilities whereas the forward algorithm takes the sum.

Each cell of the trellis in the Viterbi algorithm, $v_t(j)$, represents the probability that the model is in state j after seeing the first t observations. It also passes through the most probable state sequence $s_1, s_2, ..., s_{t1}$, given the model $\lambda = (A, B, \pi)$. The value of each cell $v_t(j)$ is computed by recursively taking the most probable path that could lead us up to that exactly cell. The probability of each cell is given by

$$v_t(j) = \max_{s_1 ... s_{t-1}} P(s_1 ... s_{t-1}, x_0, x_1, ..., x_t, s_t = j \mid \lambda) \tag{2.28}$$

The most probable path is represented by taking the maximum over all possible previous state sequences. Given that we had already computed the probability of being in every state at time t 1, we compute the Viterbi probability by taking the most probable of the extensions of the paths that lead to the current cell. For a given state s_j at time t, the value $v_t(j)$ is computed as

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(x_t) \tag{2.29}$$

Equation (2.26) extends the previous paths to compute the Viterbi algorithm and consists of three factors. That is, $v_{t-1}(i)$ the previous Viterbi path probability from the previous time step, $a_i j$ the transition probability from previous state s_i to current state s_j and $b_j(x_t)$ the state observation likelihood of the observation symbol x_t given the current state j.

One additional aspect that makes the Viterbi algorithm differ from the forward algorithm is back-pointers. In contrast to the forward algorithm that needs to produce an observation likelihood, the Viterbi algorithm must produce a probability and also the most likely state sequence. This is computed by taking best state sequence by keeping track of the path of hidden states that led to each state, and then at the end, trace the best path back to the beginning.

The learning problem: *we have a sequence of observations $x_0, ..., x_N$ and wish to know how to adjust the parameters in order to maximize the model.*

The most common way to obtain the solution to the learning problem is by using the forward-backward algorithm or the Baum-Welch algorithm, a special case of the EM-algorithm. The algorithm will let us train both the transition probabilities A and the emission probabilities B of the HMM. To learn the HMM model, we need to know what states we are in to best explain the observations. Given that the HMM parameters are fixed we can apply the forward and backward algorithm to calculate α and β from the observations. When multiplying α and β and then normalize the multiplication, we obtain the probability of state i at time t given all the observations and the model, this is called the occupation probability γ.

$$\gamma_t(j) = P(s_t = j \mid X, \lambda) = \frac{P(s_t = j, X \mid \lambda)}{P(X \mid \lambda)} = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{2.30}$$

We also need the transition probability ξ, that is, the probability of transiting from state i at time t to state j at time (t+1) given all the observations and the model. This can be computed by α and β similarly

$$\begin{aligned}\xi_t(i,j) = P(s_t = i, s_{t+1} = j \mid X, \lambda) &= \frac{P(s_t = i, s_{t+1} = j, X \mid \lambda)}{P(X \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(x_{t+1}) b_{t+1}(j)}{\sum_{t=1}^{T} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(x_t + 1) b_{t+1}(j)}\end{aligned} \tag{2.31}$$

Now it is possible to describe the Baum-Welch learning process, where parameters of the HMM is updated in such a way to maximize the quantity $P(X \mid \lambda)$. Given the starting model $\lambda = (A, B, \pi)$, we can calculate the α's and β's using equations (2.25) and (2.27) respectively and then calculate ξ's and γ's using equations (2.30) and (2.31) respectively. Next step is to update the HMM parameters according to equations (2.32) and (2.33) known as the re-estimation formulas and these are given by

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2.32}$$

and

$$\hat{b_j}(k) = \frac{\sum_{t=1 s.t. x_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.33}$$

where $\sum_{t=1 s.t. x_t = v_k}^{T} \gamma_t(j)$ means the sum over all t for which the observation at time t was a given symbol v_k from the observation vocabulary. We fix one set of parameters to improve others and continue the iteration until the solution converges.

Chapter 3

Hidden Semi-Markov model

Applications related to Financial Econometrics like risk measurement, pricing of derivatives, margin setting, and many other financial indicators rely on a suitable modelling of the distributional and temporal properties of the daily return series of stocks, indices or other assets. The unpredictable behaviour of timeseries makes it difficult the accurately modelling the properties of financial returns. As said earlier, the stock returns of a timeseries can be divided into so-called market-regimes. These various periods lead to adjustments of the asset returns via shift in their means, variances, autocorrelation, excess kurtosis, heteroskedasticity as well as skewed returns, which impact the effectiveness of time series methods that rely on stationarity. Just as the HMM, the HSMM been widely applied in financial fields due to the features in describing these complex systems of financial data analysis, by allowing to measure components distribution with several underlying components capturing the underlying regimes in the data and by inference on the "hidden" generative state processes via "noisy" indirect observations correlated to these state processes.

Rydén et al. [1998] show that a HMM mixing normal variables according to the states of an unobserved Markov chain reproduces most of the stylized facts for daily return series. However, the analysis of Rydén et al. [1998], also illustrates that the stylized fact of the very slowly decaying autocorrelation for absolute, or squared, returns cannot be described by a HMM. The lack of flexibility of a HMM to model the temporal higher order dependence can be explained by the implicit geometric distributed sojourn time in the hidden states. As an extension of the HMM, the sojourn time distribution in the HSMM can be explicitly specified by any distribution, either nonparametric or parametric, facilitating the modelling for the stylised features of stock returns. Bulla [2013] show that slow decay in the autocorrelation function can be described much better by means of HSMM's, while all other stylized facts are equally well or better reproduced. The theory of this chapter is based on the following articles: Guédon [2003], Bulla [2006], Bulla and Bulla [2006], Bulla [2013], Bulla [2011], Maruotti et al. [2019], O'Connell and Højsgaard [2011], Narimatsu and Kasai [2019], Cartella et al. [2014], Suda and Spiteri [2019], O'Connell et al. [2011], Murphy [2012], Zucchini et al. [2016] and Yu [2010]. The texts gives a thorough explanation on Hidden semi-Markov models, their inference and their applications, as well as further for reading.

To better understand the Hidden semi-Markov model, let me first introduce the semi-Markov model. Let $\{Y_t\}$ be a homogeneous Markov chain on $(1, 2, ..., K)$, with its transition probability matrix having the special feature that all its diagonal elements are zero. So in a realization of $\{Y_t\}$, no two successive values Y_t, Y_{t+1} are equal. Now allow $\{Yt\}$, plus a set of sojourn time distributions d_i on the positive integers, to generate a new process $\{S_t\}$, also on $(1, 2, ..., K)$. Each Y_t gives rise to a run of 'S-values' all equal to Y_t. The length of the run is a realization of the corresponding sojourn time distribution; that is, if $Y_t = i$, the distribution is d_i. All sojourn times are independent of each other and of earlier values of Y_t.

For example if K = 3 and $\{Yt\}$ begins with 1, 2, 1, 3, 2, a possible realization of $\{S_t\}$ is as follows:

$$111 \mid 2 \mid 1111 \mid 33 \mid 22222 \mid ...$$

Here the sequences 111 and 1111 arise from two (independent) realizations of d_1, 2 and 2222 from two realizations of d_2, and 33 from d_3. The resulting process $\{S_t\}$, which is not in general a Markov process, is called a semi-Markov process. The probabilities of self-transition in the process $\{S_t\}$, which determine the times spent in the states, are now implied by the distributions d_i. In the special case in which all the distributions d_i are geometric, $\{S_t\}$ will be a Markov process; an HMM is therefore a special case of an HSMM.

A Hidden semi-Markov model (HSMM) is an extension of the HMM by allowing the underlying process to be a semi-Markov chain, which means that for each state there is a variable duration or a sojourn time. The duration d of a given state, that is, the time spent in each state which can be seen as probabilities of self-transitions, is explicitly defined in the HSMM framework. Due to the non-zero probability of self-transition

of a non-absorbing state, the state duration of an HMM is implicitly a geometric distribution. This makes the HMM to have limitations in some applications. The HSMM allows one to utilize more flexible sojourn time distribution. In the HMM only one observation is assumed to be produced by each state, whereas for the HSMM each state can emit a sequence of observations. The number of observations produced while in state i is determined by the duration d for that state.

A HSMM consists of a pair of discrete-time stochastic processes $\{S_t\}$, called the unobserved process, and $\{X_t\}$, the observed process. Just like the HMM, the observed process $\{X_t\}$ is related to the unobserved semi-Markovian state process $\{S_t\}$ by the so-called conditional distributions. Let $X_t = \{X_1, \ldots, X_T\}$ denote the observed sequence of length T, let $S_t = \{S_1, \ldots, S_T\}$ denote the sequence of unobserved variables and let θ denote the set of model parameters.

Let $\pi_k = P(S_1 = k)$ be the initial probability of state k, $k = 1, ..., K$. $\sum_{k=1}^{K} \pi_k = 1$. For states $j, k \in \{1, ..., K\}$ with $j \neq k$, the transition probabilities are given by

$$P_{jk} = P(S_t = k \mid S_t \neq j, S_{t-1} = j), \tag{3.1}$$

satisfying $\sum_{k=1}^{K} P_{jk} = 1$ and $P_{jj} = 0$.

The diagonal elements of the transition probability matrix (TPM) of a HSMM, which is the sojourn time of each state, are required to be zero. This is because we separately model the duration of each state and do not consider the case of absorbing states. The sojourn time distribution, which is associated with each state, models how long the duration the process $\{S_t\}$ is in state j and is defined by

$$d_i(u) = P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ..., u-2 \mid S_{t+1} = j, S_t \neq j)$$

The sojourn time in the last visited state is subject to a right-censoring and is modeled by the survivor function.

$$D_j(u) = \sum_{v \leq u} d_j(v) \tag{3.2}$$

The survivor function is the key to extend the original algorithms that say there is a change of state immediately after the last observation.

The semi-Markovian state process $\{S_t\}$ of a HSMM differs from the state process of a HMM by not having the Markov property at each time t, the HSMM has the Markovian property only at the times of state changes. For the underlying semi-Markovian process, the change to a future hidden state depends on both the current state and the time spent on this state.

The observed process $\{X_t\}$ at time t is related to the state process $\{S_t\}$ by the conditional distributions and represents the state-dependent process of the HMM and fulfils the conditional independence property.

$$\begin{aligned} &P(X_t = x_t \mid X_1 = x_1, ..., X_{t-t} = x_{t-1}, S_1 = s_1, ..., S_t = s_t) \\ &\qquad = \mathrm{P}(\mathrm{X}_t = x_t \mid S_t = s_t) \end{aligned}$$

This is called the emission probability and means that the output process at time t only depends on the value of s_t. The probability that you will see x_t given that at the same time you are in state s_t. In contrast to the HMM where each state only emits one observation, the different states in the HSMM framework can emit several observation per state.

For a given length T of observation sequence $(x_1, ..., x_T)$ the length $K \leq T$ of the state sequence $(S_1, ..., S_K)$ is a random variable due to the variable state durations. We do not know in advance how to divide the observation sequence into K segments corresponding to the state sequence, and let x_t be conditioned on which state variable. A general HSMM is shown in the Figure 3.1 . In the figure, the actual sequence of events is taken to be:

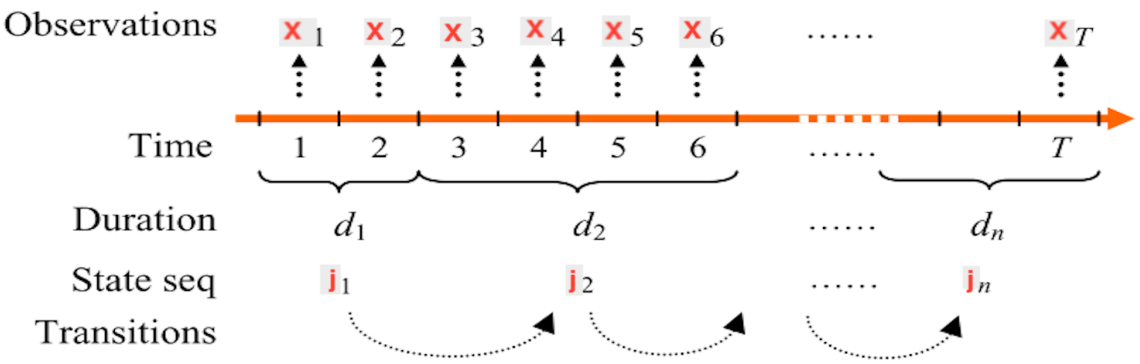


Figure 3.1: A General HSMM.

1. The first state j_1 and its duration d_1 are selected according to the state transition probability $a_{(j_0,d_0)(j_1,d_1)}$, where $a_{(j_0,d_0)}$ is the initial state and duration. State j_1 lasts for $d_1 = 2$ time units in this instance.
2. It produces two observations (x_0, x_1) according to the emission probability $b_{j_1,d_1}(x_1, x_2)$
3. It transits, according to the state transition probability $a_{(j_1,d_1)(j_2,d_2)}$, to state j_2 with duration d_2.
4. State j_2 lasts for $d_2 = 4$ time units in this instance, which produces four observations (x_3, x_4, x_5, x_6) according to the emission probability $b_{j_2,d_2}(x_3, x_4, x_5, x_6)$.
5. (j_2, d_2) then transits to (j_3, d_3), and then to (j_4, d_4) until the final observation x_T is produced. The last state j_K lasts for d_K time units, where $\sum_{k=1}^{K} d_k = T$ and T is the total number of observations

3.1 Computational aspects and inference

As a habit for mixture models, we adopt the EM-algorithm to find the maximum likelihood estimates for the parameters of our model on the basis of the observed time series $(x_1, ..., x_T)$. Once the number of latent states K has been assigned/fixed, the algorithm basically works on the complete-data likelihood. All the derivation of equations and theory in this chapter is based on the articles: Guédon [2003], Bulla [2006] and Bulla and Bulla [2006].

We will denote the observed sequence of length , $X_0 = x_0$,..., $x_{\tau-1} = X_{\tau-1}$ by $X_0^{\tau-1} = x_0^{\tau-1}$. The same convention is used for the state sequence S_t. The complete set of parameters in the model is denoted by θ. The likelihood of the complete data, i.e., the observations $X_0^{\tau-1}$ as well as the unobserved sequence $s_0^{\tau-1+u}$, is given by

$$\begin{aligned} &L_C\Big(s_0^{\tau-1+u}, X_0^{\tau-1} \mid \theta\Big) = \\ &P\Big(S_0^{\tau-1} = s_0^{\tau-1}, s_{\tau-1+v} = s_{\tau-1}, v = 1, ..., u-1, s_{\tau-1+u} \neq s_{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1} \mid \theta\Big) \end{aligned} \tag{3.3}$$

The last visited state is left at time $\tau - 1 + u$ and therefore the completed state sequence stops at this time, instead of $\tau - 1$ for algorithms without right-censoring. The contribution of the state sequence to the complete-data likelihood is given by

$$\pi_{\tilde{s_0}} d_{\tilde{s_0}}(u_0) \prod_{r=1}^{R} p_{S_{r-1}^{\sim} d_{\tilde{s_r}}}(u_r) I\Big(\sum_{r=0}^{R-1} u_r < \tau \leq \sum_{r=0}^{R} u_r\Big) \tag{3.4}$$

where $\tilde{s_0}, ..., \tilde{s_r}$ denote the $R + 1$ states visited by $s_0^{\tau-1+u}$. Combining Eqs. (3.3) and (3.4), the likelihood of the observed sequence can be calculated by summing the complete-data likelihood over all admissible paths. Compared to the classical likelihood a hidden Markov model, the likelihood function here involves an additional sum over all possible prolongations of the state sequence $s_0^{\tau-1}$ up to the exit from the last visited state:

$$L(\theta) = \sum_{s_0,...,s_{\tau-1}} \sum_{\tau+} L_C\Big(s_0^{\tau-1+u}, X_0^{\tau-1} \mid \theta\Big) \tag{3.5}$$

where $\sum_{s_0,...,s_{\tau-1}}$ denotes the summation over every possible state sequence of length τ, and $\sum_{\tau+}$ denotes the sum over every supplementary duration from time spent in the state occupied at time $\tau - 1$. Instead of the successively visited states, the sojourn times and the outputs emitted in these states, only the outputs are

observed. Hence, we are faced with an incomplete-data problem and the EM algorithm is a natural candidate for deriving the nonparametric maximum likelihood estimator

Another algorithm that is widely used for the maximization of the log likelihood of a HSMM model is the direct numerical maximization (DNM). The difficulty of the maximization of the likelihood of a HSMM model varies with the component distributions chosen and may in some cases require numerical maximization methods when an explicit solution is not available, e.g. for the t distribution in Appendix B.2. Direct numerical maximization (DNM) has some appealing properties, especially concerning the treatment of missing observations, flexibility in fitting complex models and the speed of convergence in the neighbourhood of a maximum. The main disadvantage of this method is its relatively small circle of convergence.

In R, there exists some integrated functions that can be used to carry out DNM and minimize the negative log-likelihood, namely the nlm() and the optim(). The function nlm() minimize the negative log-likelihood using a Newton-type algorithm, while the function optim() minimize the negative log-likelihood using the Nelder-Mead simplex algorithm. Bulla [2006] he said that, in general, the Nelder-Mead algorithm is more stable; however, it may also get stuck in local minima and is rather slow when compared to the Newton-type minimization. Direct numerical maximization of the likelihood using Newton-type algorithms generally converges faster than the EM algorithm, especially in the neighbourhood of a maximum. However, it requires more accurate initial values than the EM to converge at all.

Another type of algorithm that can be used is the so-called hybrid algortihm, which is a combination of the EM-algortihm and the DNM. While the EM-algorithm has a slow E-step and a complicated M-step, it is more stable than the DNM and has a large circle of convergence, and in addition, doesn't need as accurate starting values as the DNM. Hybrid algorithms, which are constructed by combining the EM algorithm with a rapid algorithm with strong local convergence, using e.g. a Newton-type algorithm, yields the stability and tha large circle of convergence from the EM algorithm along with superlinear convergence of the Newton-type algorithm in the neighbourhood of the maximum. However, in this paper, we are going to focus on the EM-algorithm.

3.1.1 EM-algorithm

The likelihood of a HSMM given in Eq. (3.5) could be easily evaluated if the state sequence was known. However, this is not the case and we are hence confronted with a missing data problem. A popular method of dealing with this type of problem is the expectation maximization (EM) algorithm, an iterative procedure which increases the likelihood monotonically until it reaches a stationary point. After assigning initial values to the parameters, the EM algorithm is implemented by successively iterating the following two steps:
E-step: Compute the Q-function

$$Q\big(\theta, \theta^{(k)}\big) = E\Big[L_C\big(s_0^{\tau-1+u}, x_0^{\tau-1} \mid \theta\big) \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big] \tag{3.6}$$

the conditional expectation of the complete-data log-likelihood, where $\theta^{(k)}$ denotes the current estimate of the parameter vector θ.
M-step: Compute $\theta^{(k+1)}$, the parameter values that maximize the function Q w.r.t. θ, i.e.,

$$\theta^{(k+1)} = \text{argmax}_\theta Q\big(\theta, \theta^{(k)}\big) \tag{3.7}$$

The two steps are repeated until a stationary point is reached. In our case, the EM algorithm maximizes $L(\theta)$ from Eq. (3.5). Each iteration of the EM algorithm increases $L()$ and, generally, the sequence of re-estimated parameters $\theta^{(k)}$ converge to a local maximum of $L()$

The main difficulty of the EM algorithm is the E-step. To obtain a mathematically tractable formulation of the Q-function, the conditional expectation has to be rewritten pathwise. The conditional distribution of the missing observations is given by:

$$P\Big(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta\Big)$$

and the distribution of the complete-data by

$$P\Big(S_0^{\tau-1+u} = s_0^{\tau-1+u}, X_0^{\tau-1} = x_0^{\tau-1} \mid \theta\Big).$$

Hence, the first step is the transformation of the E-step equation (3.6) into

$$Q\big(\theta,\theta^{(k)}\big)=\sum_{s_0,\ldots,s_{\tau-1}}\sum_{\tau+}L_C\Big(s_0^{\tau-1+u},x_0^{\tau-1}\mid\theta\Big)P\Big(S_0^{\tau-1+u}=s_0^{\tau-1+u}\mid X_0^{\tau-1}=x_0^{\tau-1},\theta\Big). \tag{3.8}$$

Secondly, consider the contribution of a specific path to the likelihood of a HSMM given by Eq.(3.4). Adding the contribution of the observed sequence, the complete-data likelihood becomes

$$L_C\Big(s_0^{\tau-1+u},x_0^{\tau-1}\mid\theta\Big)=\pi_{\tilde{s_0}}d_{\tilde{s_0}}(u_0)\prod_{k=1}^{K}p_{S_{r-1}^{\sim}}d_{\tilde{s_r}}(u_r)\prod_{t=0}^{\tau-1}b_{s_t}(x_t) \tag{3.9}$$

Taking the logarithm splits Eq. (3.9) into four independent terms and Eq. (3.8) becomes

$$\begin{aligned}Q\big(\theta,\theta^{(k)}\big)=&\sum_{s_0,\ldots,s_{\tau-1}}\sum_{\tau+}\Big[log\pi_{\tilde{s_0}}+\Big(\sum_{k=1}^{K}logp_{S_{r-1}^{\sim}}\Big)+\Big(\sum_{k=0}^{K}d_{\tilde{s_r}}(u_r)\Big)+\Big(\sum_{t=0}^{\tau-1}b_{s_t}(x_t)\Big)\Big]\\&\cdot P\Big(S_0^{\tau-1+u}=s_0^{\tau-1+u}\mid X_0^{\tau-1}=x_0^{\tau-1},\theta\Big).\end{aligned} \tag{3.10}$$

The four terms in Eq. (3.10) correspond to the initial, the transition, the sojourn time and the observation probabilities. In a last step, the summation over all paths is marginalized, and the four terms can be written as:

$$\sum_{j=0}^{J-1}P\Big(S_0=j\mid X_0^{\tau-1}=x_0^{\tau-1}\mid\theta^{(k)}\Big)\log\pi_j \tag{3.11}$$

$$\sum_{i=0}^{J-1}\sum_{i\neq j}\sum_{t=0}^{\tau-2}P(S_t=j,S_t=i\mid X_0^{\tau-1}=x_0^{\tau-1}\mid\theta^{(k)}\Big)\log p_{ij} \tag{3.12}$$

$$\begin{aligned}\sum_{u}\Big[\sum_{t=0}^{\tau-2}&P\Big(S_{t+u+1}\neq j,S_{t+u-v}=j,v=0,\ldots,u-1,\neq j\mid X_0^{\tau-1}=x_0^{\tau-1}\mid\theta^{(k)}\Big)\\&+P\Big(S_u\neq j,S_{u-v}=j,v=1,\ldots,u\mid X_0^{\tau-1}=x_0^{\tau-1}\mid\theta^{(k)}\Big)\Big]\log d_j(u)\end{aligned} \tag{3.13}$$

$$\sum_{j=0}^{J-1}\sum_{t=0}^{\tau-1}P\Big(S_t=j\mid X_0^{\tau-1}=x_0^{\tau-1}\mid\theta^{(k)}\Big)\log b_j(x_t) \tag{3.14}$$

The re-estimation formulas for these four quantities can be obtained by maximizing each of the terms separately. The, Eqs. (3.11), (3.12) and (3.14) can be calculated via the dynamic programming method known as the forward-backward algorithm. As concerns the updating of Eq. (3.13) Guédon [2003] provides a version of the forward-backward algorithm which is implemented in the mhsmm package O'Connell and Højsgaard [2011] for R.

3.1.2 Forward-Backward Algorithm

The implementation of the E-step of the EM algorithm is performed by the forward-backward algorithm. It computes all the re-estimation quantities for all times t and for all states j. The M-step maximizes each of the terms w.r.t. θ to obtain the next set of initial values for the E-step of the following iteration.

In the case of a hidden Semi-Markov model, the forward-backward algorithm is based on the following decomposition

$$\begin{aligned}L1_j(t)&=P(S_{t+1}\neq j,S_t=j\mid X_0^{\tau-1}=x_0^{\tau-1})\\&=\frac{P(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1}\mid S_{t+1}\neq j,s_t=j)}{P(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1}\mid X_0^{\tau-1}=x_0^{\tau-1})}P(S_{t+1}\neq j,S_t=j\mid X_0^{t}=x_0^{t})\\&=B_j(t)F_j(t)\end{aligned} \tag{3.15}$$

which expresses the conditional independence between the past and the future of the process at the times of a state change of the hidden semi-Markov chain. It forms the basis of the forward-backward algorithm for HSMM's. In the case of a hidden Markov chain, decomposition (2.31) naturally fits the EM estimate

requirements while, in the case of a hidden semi-Markov chain, decomposition (3.15) does not directly fit the EM estimate requirements.

Guédon [2003] developed a powerful estimation algorithm based on the decomposition (3.15). The recursion's complexity both in time and in space is similar to that of the forward recursion, which is $O(J\tau(J+\tau))$-time (in the worst case) and O $(J\tau)$-space. This means that the computation of $L_j(t) = P(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})$, which is the forward-backward algorithm for HMM's, as well as $L1_j(t) = P(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})$ require the same complexity. This allows one to fit even long sequences of observations in a reasonable amount of time. Moreover he relaxes the assumption that the last visited state terminates at the time of the last observation.

The Forward Iteration

The forward iteration involves the computation of the forward-probabilities $F_j(t) = P(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t)$ for each state j forward from time 0 to time $\tau - 1$ and can be presented as follows.

The start of the loop at t = 0 can be simplified to

$$\begin{aligned} F_j &= P(S_1 \neq j, S_0 = J \mid X_0 = x_0) \\ &= \pi_j d_j(1), \end{aligned} \tag{3.16}$$

and

$$\begin{aligned} F_j(t) &= P(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t) \\ &= \frac{b_j(x_t)}{N_t}\left[\sum_{u=1}^{t}\left(\prod_{v=1}^{u-1}\frac{b_j(x_{t-v})}{N_{t-v}}\right)d_j(u)\sum_{i\neq j}p_{ij}F_i(t-u)\right. \\ &\left. + \left(\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right)d_j(t+1)\pi_j\right], \end{aligned} \tag{3.17}$$

for all $t \in \{0, ..., \tau - 2\}$ and $j \in \{0, ..., J-1\}$, where $N_t = P(X_t = x_t \mid X_0^{t-1} = x_0^{t-1})$ is a normalizing factor. For time $\tau - 1$ and $j \in \{0, ..., J-1\}$, the forward iteration can be rewritten as:

$$\begin{aligned} F_j(t) &= P(S_{\tau-1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &= \frac{b_j(x_{\tau-1})}{N_{\tau-1}}\left[\sum_{u=1}^{\tau-1}\left(\prod_{v=1}^{u-1}\frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}}\right)D_j(u)\sum_{i\neq j}p_{ij}F_i(\tau-1-u)\right. \\ &\left. + \left(\prod_{v=1}^{\tau-1}\frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}}\right)D_j(\tau)\pi_j\right], \end{aligned} \tag{3.18}$$

The exact time spent in the last visited state is unknown, only the minimum time spent in this state is known. Therefore, the probability mass functions of the sojourn times in state j of the general forward recursion formula (3.17) are replaced by the corresponding survivor functions in (3.2).

The normalizing factor N_t is directly obtained during the forward recursion. We obtain, $t \in \{0, ..., \tau - 1\}$:

$$\begin{aligned} N_j(t) &= P(X_t = x_t \mid X_0^{t-1} = x_0^{t-1}) \\ &= \sum_j P(S_t = j, X_t = x_t \mid X_0^{t-1} = x_0^{t-1}) \\ &= \sum_j b_j(x_t) N_t\left[\sum_{u=1}^{t}\left(\prod_{v=1}^{u-1}\frac{b_j(x_{t-v})}{N_{t-v}}\right)D_j(u)\sum_{i\neq j}p_{ij}F_i(t-u) + \left(\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right)D_j(t+1)\pi_j\right], \end{aligned} \tag{3.19}$$

The Backward Iteration

The backward iteration consists of computing $L_j(t) = P(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})$ for each state j backward from time $\tau - 1$ to time 0. The backward iteration is initialized for $t = \tau - 1$ by, $j = (0, ..., J-1)$:

$$L_j(\tau - 1) = P(S_{\tau-1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}) = F_j(\tau - 1)$$

The key point in this step lies in rewriting the quantity $L_j(t)$ as a sum of three terms:

$$\begin{aligned} L_j(t) &= P(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &= P(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &+ P(S_{t+1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &- P(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &= L1_j(t) + L_j(t+1) - P(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}) \end{aligned} \tag{3.20}$$

The backward recursion is based on $L1_j(t)$ for $t \in \{\tau-2, ..., 0\}$ and $J \in \{0, ..., J-1\}$

$$\begin{aligned} L1_j(t) = \Bigg[\sum_{k \neq j} \Bigg[\sum_{u=1}^{\tau-2-t} \frac{L1_k(t+u)}{F_k(t+u)} \Bigg(\prod_{v=0}^{u-1} \frac{b_k(x_{t+u-v})}{N_{t+u-v}} d_k(u) \Bigg) \\ + \Bigg(\prod_{v=0}^{\tau-2-t} \frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}} \Bigg) D_k(\tau-1-t) \Bigg] p_{jk} \Bigg] F_j(t) \end{aligned} \tag{3.21}$$

The third term in (3.20), for for $t \in \{\tau-2, ..., 0\}$ and $J \in \{0, ..., J-1\}$, is given by

$$\begin{aligned} &P(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1}) \\ &= \Bigg[\sum_{u=1}^{\tau-2-t} \frac{L1_j(t+u)}{F_j(t+u)} \Bigg(\prod_{v=0}^{u-1} \frac{b_j(x_{t+u-v})}{N_{t+u-v}} d_j(u) \Bigg) \\ &+ \Bigg(\prod_{v=0}^{\tau-2-t} \frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}} \Bigg) D_j(\tau-1-t) \Bigg] \sum_{i \neq j} p_{ij} F_i(t) \end{aligned} \tag{3.22}$$

The computation of $L_j(t)$ may appear at first sight relatively intricate but, in fact, the computations of $L1_j(t) = P(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})$ in (3.21) and $P(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1})$ in (3.22) may easily be performed by introducing the following auxiliary quantities:

$$G_j(t+1, u) = \frac{L1_j(t+u)}{F_j(t+u)} \Bigg(\prod_{v=0}^{u-1} \frac{b_j(x_{t+u-v})}{N_{t+u-v}} \Bigg) d_j(u), u = 1, ..., \tau-2-t, \tag{3.23}$$

$$G_j(t+1, \tau-1-t) = \Bigg(\prod_{v=0}^{\tau-2-t} \frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}} \Bigg) D_j(\tau-1-t) \tag{3.24}$$

and

$$G_j(t+1) = \frac{P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid S_{t+1} = j, S_t \neq j)}{P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid X_0^t = x_0^t)} \tag{3.25}$$

At each time t, these auxiliary quantities should be precomputed. Then,

$$L1_j(t) = \Bigg[\sum_{k \neq j} G_k(t+1) p_{jk} \Bigg] F_j(t) \tag{3.26}$$

and

$$\begin{aligned} &P(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}) \\ &= \frac{P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid S_{t+1} = j, S_t \neq j)}{P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid X_0^t = x_0^t)} P(S_{t+1} = j, S_t \neq j \mid X_0^t) = x_0^t \\ &= G_j(t+1) \sum_{1 \neq j} p_{ij} F_i(t) \end{aligned} \tag{3.27}$$

Because, for each $t < \tau - 1$, $L1_j(t) = B_j(t) F_j(t)$, the backward recursion based on $B_j(t)$ is directly deduced from (3.21)

$$
\begin{aligned}
B_j(t) = \sum_{k \neq j} \Bigg[\sum_{u=1}^{\tau-2-t} B_k(t+u) \Big(\prod_{v=0}^{u-1} \frac{b_j(x_{t+u-v})}{N_{t+u-v}} d_k(u) \Big) \\
+ \Big(\prod_{v=0}^{\tau-2-t} \frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}} \Big) D_k(\tau-1-t) \Bigg] p_{jk}
\end{aligned}
\tag{3.28}
$$

3.1.3 Parameter re-estimation

The second step of the EM-algorithm consists of a re-estimation of the parameters of the hidden semi-Markov model. This step determines the likelihood-increasing next set of parameters θ^{k+1} by:

$$\theta^{(k+1)} = \text{argmax}_\theta Q\big(\theta, \theta^{(k)}\big)$$

The re-estimation of the parameters is obtained by maximizing the different terms of the M-step, $Q(\theta \mid \theta^k$. We showed that $Q(\theta \mid \theta^k$ could be decomposed into four terms, each term depending on a given subset of θ. In the following, we derive the re-estimation formula for each parameter subset by maximizing the terms (3.11), (3.12), (3.13) and (3.14). The re-estimation formula for the initial parameters is given by

$$\pi_j^{k+1} = P(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}) = L_j(0) \tag{3.29}$$

The re-estimation formula for the transition probabilities can be written as

$$
\begin{aligned}
P_{ij}^{k+1} &= \frac{\sum_{t=0}^{\tau-2} P(S_{t+1} = j, S_t = i \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}}{\sum_{t=0}^{\tau-2} P(S_{t+1} \neq i, S_t = i \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}} \\
&= \frac{\sum_{t=0}^{\tau-2} G_j(t+1) p_{ij} F_i(t)}{\sum_{t=0}^{\tau-2} L1_i(t)}
\end{aligned}
\tag{3.30}
$$

The numerator in equation (3.30) is directly extracted from the computation of $L1_j(t)$, equation (3.26). Concerning the state occupancy distribution, for each non-absorbing state j we have:

$$
\begin{aligned}
& Q_d \Big[(d_j(u)) \mid \theta^{(k)} \Big] \\
&= \sum_u \Bigg[\sum_{t=0}^{\tau-2} P\Big(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ..., u-1, \neq j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)} \Big) \\
&+ P\Big(S_u \neq j, S_{u-v} = j, v = 1, ..., u \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)} \Big) \Bigg] \log d_j(u) \\
&= \sum_u \eta_{j,u}^{(k)} \log d_j(u)
\end{aligned}
\tag{3.31}
$$

For the second equality in term (3.31), $u \leq \tau - 2 - t$ is directly extracted from the computation of $L_j(t)$

$$
\begin{aligned}
& P\Big(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ..., u-1, \neq j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)} \Big) \\
&= G_j(t+1, u) = \sum_{i \neq j} p_{ij} F_i(t)
\end{aligned}
$$

while for $u > \tau - 2 - t$, we obtain

$$
\begin{aligned}
& P\Big(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ..., u-1, \neq j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)} \Big) \\
&= \Big(\prod_{v=0}^{\tau-2-t} \frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}} \Big) d_j(u) \sum_{i \neq j} p_{ij} F_i(t)
\end{aligned}
$$

The term in (3.31) corresponding to the time spent in the initial state requires some supplementary computation at time t = 0,
$u \leq \tau - 1$

$$P\Big(S_u \neq j, S_{u-v} = j, v = 1, ..., u \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)$$
$$= \frac{L1_j(u-1)}{F_j(u-1)}\Big(\prod_{v=1}^{u} \frac{b_j(x_{u-v})}{N_{u-v}}\Big) d_j(u)\pi_j$$

$u > \tau - 1$

$$P\Big(S_u \neq j, S_{u-v} = j, v = 1, ..., u \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)$$
$$= \Big(\prod_{v=1}^{\tau} \frac{b_j(x_{\tau-v})}{N_{\tau-v}}\Big) d_j(u)\pi_j$$

By noting that the former computation can merge into

$$\sum_u \Big[\sum_{t=0}^{\tau-2} P\Big(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ..., u-1, \neq j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)$$
$$+ P\Big(S_u \neq j, S_{u-v} = j, v = 1, ..., u \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)\Big]$$
$$= \sum_{t=0}^{\tau-2} P\Big(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big) + P\Big(S_{\tau-1} = j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)$$
$$= \sum_{t=0}^{\tau-2} L1_j(t) + L_j(\tau - 1)$$

the re-estimated state occupancy probabilities are then given by

$$d_j^{(k+1)}(u) = \frac{\eta_{j,u}^{(k)}}{\sum_v \eta_{j,v}^{(k)}} = \frac{\eta_{j,u}^{(k)}}{\sum_{t=0}^{\tau-2} L1_j(t) + L_j(\tau - 1)} \tag{3.32}$$

The re-estimated observation probabilities are given by

$$\begin{aligned} b_j^{k+1}(y) &= \frac{\sum_{t=0}^{\tau-1} P\Big(X_t = y, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)}{\sum_{t=0}^{\tau-1} P\Big(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1} \mid \theta^{(k)}\Big)} \\ &= \frac{\sum_{t=0}^{\tau-1} L_j(t) I(x_t = y)}{\sum_{t=0}^{\tau-1} L_j(t)} \end{aligned} \tag{3.33}$$

All the quantities involved in the re-estimation formulas (3.29) (3.30) (3.32) (3.33) are directly extracted from the backward recursion. There are only a few additional computation that should be noted, the computation concerning the contributions at time t=0 and the contributions of the time spent in the last visited state to the re-estimation quantities of the state occupancy distributions.

3.1.4 Viterbi Algorithm

Given a model, we are interested in the most likely sequence of states, given the sequence of observations. That is, we wish to find the sequence of states that maximizes $P(S \mid X, \theta)$. Calculating $P(S \mid X, \theta)$ for every possible state sequence is not computationally feasible. A dynamic programming technique known as the Viterbi algorithm, can be used to maximize $P(S \mid X, \theta)$. Because the state process is a semi-Markov chain, we have for all t

$$\begin{aligned} &\max_{s_0,...,s_{\tau-1}; s_{t+1} \neq s_t} P\Big(S_0^{\tau-1} = s_0^{\tau-1} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta\Big) \\ &= \max_{s_t}\Big(\max_{s_{t+1},...,s_{\tau-1}} P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1}, S_{t+1}^{\tau-1} = s_{t+1}^{\tau-1} \mid S_{t+1} \neq s_t, S_t = s_t) \\ &\times \max_{s_0,...,s_{t-1}} P(S_{t+1} \neq j, S_t = j \mid S_0^t = s_0^t, X_0^t = x_0^t)\Big) \end{aligned} \tag{3.34}$$

Let us define

$$\alpha_j(t) = \max_{s_0,\dots,s_{t-1}} P(S_{t+1} \neq s_t, S_t = s_t, S_0^{t-1} = s_0^{t-1}, X_0^t = x_0^t)$$

Thus, the equation (3.34) can be rewritten as

$$\begin{aligned} &\max_{s_0,\dots,s_{\tau-1};s_{t+1}\neq s_t} P\Big(S_0^{\tau-1} = s_0^{\tau-1} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta\Big) \\ &= \max_j \left(\max_{s_{t+1},\dots,s_{\tau-1}} P(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1}, S_{t+1}^{\tau-1} = s_{t+1}^{\tau-1} \mid S_{t+1} \neq j, S_t = j)\alpha_j(t) \right) \end{aligned} \tag{3.35}$$

Based on the decomposition of equation (3.35), we can build the following recursion $t = 0, ..., \tau - 2; j = 0, ..., J - 1$

$$\begin{aligned} \alpha_j(t) &= \max_{s_0,\dots,s_{t-1}} P(S_{t+1} \neq s_t, S_t = s_t, S_0^{t-1} = s_0^{t-1}, X_0^t = x_0^t) \\ &= b_j(x_t) \max \left[\max_{1\leq u\leq t} \left[\left(\prod_{v=1}^{u-1} b_j(x_{t-v}) \right) d_j(u) \max_{i\neq j}(p_{ij}\alpha_i(t-u)) \right], \right. \\ &\left. \left(\prod_{v=1}^{t} b_j(x_{t-v}) \right) d_j(t+1)\pi_j \right] \end{aligned} \tag{3.36}$$

The right censoring of the sojourn time in the last visited state makes particular the case $t = \tau - 1$ and $j = 0, ..., J - 1$

$$\begin{aligned} \alpha_j(\tau-1) &= \max_{s_0,\dots,s_{\tau-2}} P(S_{\tau-1} = j, S_0^{\tau-2} = s_0^{\tau-2}, X_0^{\tau-1} = x_0^{\tau-1}) \\ &= b_j(x_{\tau-1}) \max \left[\max_{1\leq u\leq \tau-1} \left[\left(\prod_{v=1}^{u-1} b_j(x_{\tau-1-v}) \right) D_j(u) \max_{i\neq j}(p_{ij}\alpha_i(\tau-1-u)) \right], \right. \\ &\left. \left(\prod_{v=1}^{\tau-1} b_j(x_{\tau-1-v}) \right) D_j(\tau)\pi_j \right] \end{aligned} \tag{3.37}$$

The likelihood of the optimal state sequence associated with the observed sequence $X_0^{\tau-1}$ is $\max_j \alpha_j(\tau - 1)$

Chapter 4

Distributions

In this chapter I will introduce the different distributions to be used in the HSMM framework. The distributions that I will cover in this paper is the Normal distribution, the Skew-Normal distribution, the T-distribution and the Skew-t distribution. The theory in this chapter is based on the following articles: Fama [1965], Blattberg and Gonedes [1974], Eling [2012] and Davis [2015].

4.1 Symmetric Distributions

4.1.1 The Normal Distribution

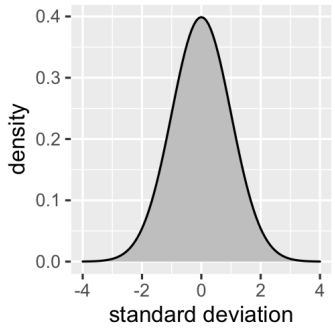


Figure 4.1: Normal distribution

A continuous random variable X has a normal distribution if its probability density function has the form:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{4.1}$$

where the parameter μ is the mean or expectation of the distribution and σ is the standard deviation.

The Normal distribution is a probability distribution with two parameters, mean and sigma, and is symmetric about the mean and shaped like a bell curve. The data near the mean are more frequent in occurrence then data far from the mean. About 68 percent of the observations are within one standard deviation away from the mean; about 95 percent of the observations lie within two standard deviations; and about 99.7 percent are within three standard deviations, so three standard deviations cover all but 0.27 percent of the data. The Normal distribution is therefore not appropriate when the model have outliers, values that lie many standard deviations away from the mean. In such cases models with heavier tails is more appropriate. Normal distribution is an appropriate choice when the observations are assumed to be normal.

In finance, the use of normal distribution is a convenient simplification because of the simplicity in the approach and to develop the basic theoretical structure. However, real life data rarely follow a perfect normal distribution. The aspect of non-normality with skewness and kurtosis of financial market returns makes the normal distribution unable to capture the essence of the data.

4.1.2 T-distribution

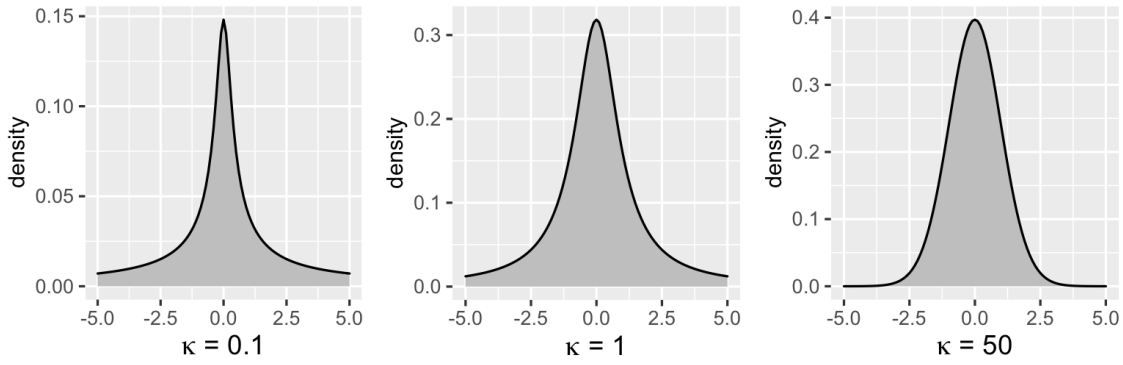


Figure 4.2: T-distribution with κ = kurtosis

The t-distribution is a symmetric and bell-shaped distribution just like the normal distribution. The difference is that the t-distribution is leptokurtic, and so has higher kurtosis than the normal distribution, which means that it has heavier tails. The probability of getting values very far from the mean is larger with a t-distribution than a normal distribution. Because the t-distribution has heavier tails than a normal distribution, it can be used as a model for financial returns that exhibit excess kurtosis. This will allow for a more realistic calculation of different risk measures, such as Value at Risk and Expected Shortfall

The kurtosis is determined by a parameter called degrees of freedom (df). Small values of df gives heavier tails, and higher values makes the t-distribution resemble a standard normal distribution.

The t-distribution has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \tag{4.2}$$

where ν is the number of degrees of freedom and Γ is the gamma function.

4.2 Skew-Elliptical Distributions

A lot of financial theory is based on the assumption that the probability distribution of asset returns is normal distributed. However, many papers indicate that this assumption is not supported, and that models based on this assumption fail to satisfy and fit real-world data. The Normal distribution, which is centred around a zero mean and with perfect symmetry, is an inadequate model for asset returns where the presence of fat tails and asymmetry has been widely reported.

Therefore, skew-elliptical distributions such as the skew-normal and the skew t-distribution might be promising since they preserve the advantages of the normal and the t-distribution and also have the benefit of flexibility with regard to skewness and kurtosis.

4.2.1 Skew-Normal Distribution

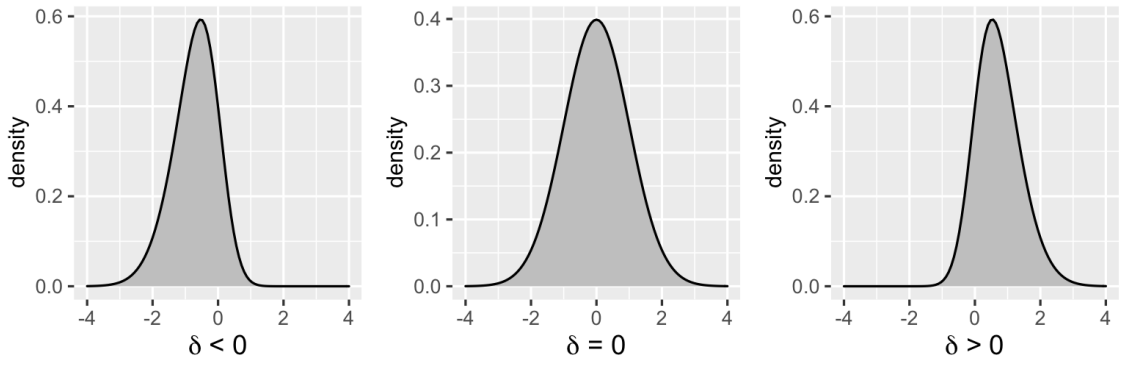


Figure 4.3: Skew-N distribution with δ = skewness

A continuous random variable X has a skew-normal distribution if its probability density function has the form:

$$f(x) = 2\phi(x)\Phi(\alpha x) \tag{4.3}$$

where $\phi(x)$ denotes the standard normal probability density function given by

$$\phi(x) = \frac{1}{\sqrt{2}\pi} e^{\frac{-x^2}{2}} \tag{4.4}$$

and $\Phi(x)$ denotes the cumulative distribution function given by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = \frac{1}{2}[1 + erf(\frac{x}{\sqrt{2}})] \tag{4.5}$$

where erf is the error function. Equation (4.3) is called the skew normal distribution with shape parameter δ, $X \sim SN(0, 1, \alpha)$. The parameter δ determines the skewness of the distribution. Higher positive values of δ implies a more right-skewed distribution, and negative values of δ implies a more left-skewed distribution. When the shape parameter, $\alpha = 0$, the skew normal distribution reduces to a standard normal distribution.

The skew normal distribution extends the normal distribution in that it can extend the range of available skewness. This makes the skew-normal distribution more appropriate when modeling data with skewness, such as asset returns. However the range of potential skewness is still limited as it only can take values of skewness from -1 to 1. The skew t-distribution, however, can take values of skewness more extreme to that of a skew-normal distribution.

4.2.2 Skew T-Distribution

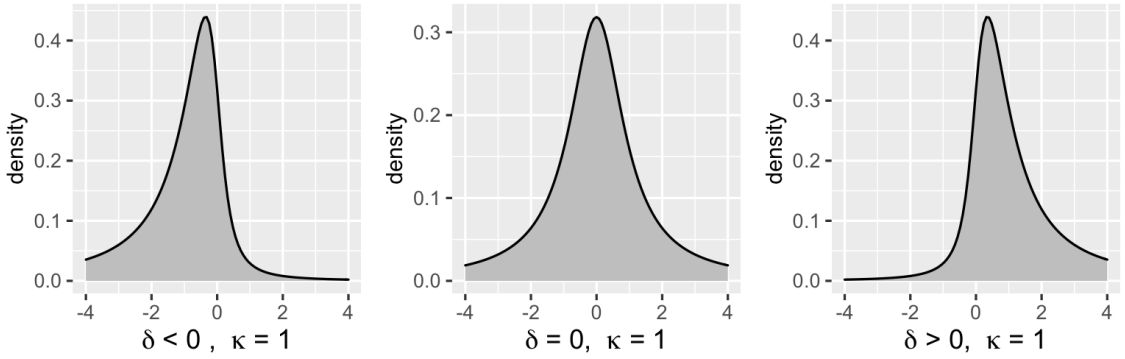


Figure 4.4: Skew-T distribution with δ = skewness, and κ = kurtosis

The skew t-distribution allows regulating both the skewness and kurtosis of a distribution. This attribute is particularly useful in empirical applications when we want to consider distributions with higher kurtosis than the normal distribution and higher skewness than the skew normal distribution, which is often the case in finance applications.

We define the standardized skewed t-distribution using the transformation

$$X = \frac{Z}{\sqrt{W/\nu}} \tag{4.6}$$

where $W \sim \chi^2(\nu)$ with ν degrees of freedom and Z is an independent $SN(0, 1, \alpha)$.

Compared to the skew normal distribution, the skew t-distribution can take more extreme values of both kurtosis and/or skewness, which makes it more adaptable when modelling data with both kurtosis and skewness present.

Chapter 5

Financial returns

In the financial world, the goal of investing is to make profit. The profit of the investment depends upon three factors, the changes in price market, the amount invested, and the amount of assets being held in the portfolio. In order to do a proper analysis of our investment data, we need to convert the price data into values, that is, return values, which further can be modelled by a statistical distribution. If we want to apply the machinery of statistical inference and the probability distribution to our return values, we must make the assumption that the returns are stationary, which means that the statistical properties of the process generating the time series does not change over time. The statistical properties of the future returns such as mean, variance, auto-correlation, etc. are all constant over time, when the stationarity assumption is present. Without this assumption the uncertainty is much harder to measure, thus Finance, is a grey zone between measurable and unmeasurable uncertainty.

In this section I will introduce some of the most common return measures in the context of financial analysis. The theory in this chapter is based on the article/textbooks Ruppert and Matteson [2015], Brockwell and Davis [2002], Palachy [2019] and Greunen [2011] which offers a thorough introduction in financial returns, stationarity and time series analysis. The texts also provide references for further reading.

5.1 Net-return

Let P_t be the price of an asset at time t. Assuming no dividends, the net return over the holding period from time $t-1$ to time t is

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 \tag{5.1}$$

where P_t is the price of the financial instrument at time t. The numerator, PtP_{t1}, is the profit during the holding period, with negative profit meaning a loss. The denominator, P_{t-1}, is the initial investment at the start of the holding period, and therefore, the net-return can be viewed as the relative revenue or profit rate. Based on the returns at time t, the aggregation of daily returns over the period k can be expressed as the product of single-period net returns. Based on the returns at time t, we can merge the daily returns over the period k and express it as the product of a single period net return. Let us consider the case of the returns to an investment made in time t until time $t+k$. In this case, we define the simple multi-period return as

$$R_t(k) = \frac{P_{t+k}}{P_t} - 1 = \frac{P_{t+1}}{P_t} \cdot \frac{P_{t+2}}{P_{t+1}} \cdots \frac{P_{t+k}}{P_{t+k-1}} - 1 = \prod_{i=1}^{k} \frac{P_{t+i}}{P_{t+i-1}} - 1 \tag{5.2}$$

5.2 Gross-return

The simple gross return is defined as

$$\frac{P_t}{P_{t-1}} = 1 + R_t \tag{5.3}$$

The gross-returns are scale-free, meaning that they do not depend on units, although they depend on the unit of time t. The gross return over k periods is the product of the k single-period gross returns from time t to time $t+k$

$$\begin{aligned} 1 + R_t(k) = \frac{P_{t+k}}{P_t} &= \frac{P_{t+1}}{P_t} \cdot \frac{P_{t+2}}{P_{t+1}} \cdots \frac{P_{t+k}}{P_{t+k-1}} \\ &= (1 + R_t) \cdot (1 + R_{t+1}) \cdots (1 + R_{t+k}) \\ &= \prod_{i=0}^{k} (1 + R_{t+i}) \end{aligned} \tag{5.4}$$

5.3 Log-returns

Log returns, also called continuously compounded returns, are denoted by r_t and defined as

$$r_t = log(1 + R_t) = log(\frac{P_t}{P_{t-1}}) = log(P_t) - log(P_{t-1}) = p_t - p_{t-1} \tag{5.5}$$

where $p_t = log(P_t)$ is called the "log-price", and $r_t = log(P_t) - log(P_{t-1})$ correspond to the log-return differences of the asset prices at time t. Since returns are smaller in magnitude over shorter periods, we can expect net-returns and log-returns to be similar for daily returns, less similar for monthly returns, and not necessarily similar for longer periods such as many years. The net-return and log-return have the same sign, though the magnitude of the log-return is smaller than that of the return if they are both positive, or larger if they are both negative. One advantage of using log-returns is simplicity of multiperiod returns. A k-period log-return is simply the sum of the single period log-returns, rather than the product as for net-returns. Let us consider the case of the log-returns of an investment made in time t until time $t + k$. In this case, we define the simple multi-period log-return as

$$\begin{aligned} r_t(k) &= log\{1 + R_t(k)\} \\ &= log\{(1 + R_t)(1 + R_{t+1}) \cdots (1 + R_{t+k})\} \\ &= log(1 + R_t) + log(1 + R_{t+1}) + \cdots + log(1 + R_{t+k}) \\ &= r_t + r_{t+1} + r_{t+k} = \sum_{i=0}^{k} r_{t+i} \end{aligned} \tag{5.6}$$

5.4 Adjustment for dividends

Many stocks, especially those of mature companies, pay dividends that must be accounted for when computing returns. Similarly, bonds pay interest. If a dividend D_t is paid prior to time t, then the net-return at time t is defined as

$$R_t = \frac{P_t + D_t}{Pt - 1} - 1 \tag{5.7}$$

and then the gross-return at time t is defined as

$$1 + R_t = \frac{P_t + D_t}{P_{t-1}} \tag{5.8}$$

and the log-return

$$r_t = log(\frac{P_t + D_t}{Pt - 1}) \tag{5.9}$$

If there is no dividend between time $t - 1$ and time t, $D_t = 0$. The multi-period net-return from time t to time $t + k$ with dividends is defined as

$$\begin{aligned} R_t(k) &= \frac{P_{t+k} + D_{t+k}}{P_t} - 1 \\ &= \frac{P_{t+1} + D_{t+1}}{P_t} \cdot \frac{P_{t+2} + D_{t+2}}{P_{t+1}} \cdots \frac{P_{t+k} + D_{t+k}}{P_{t+k-1}} - 1 \\ &= \prod_{i=1}^{k} \frac{P_{t+i} + D_{t+i}}{P_{t+i-1}} - 1 \end{aligned} \tag{5.10}$$

a k-period gross-return with dividends is defined as

$$\begin{aligned} 1 + R_t(k) = \frac{P_{t+k} + D_{t+k}}{P_t} &= \frac{P_{t+1} + D_{t+1}}{P_t} \cdot \frac{P_{t+2} + D_{t+2}}{P_{t+1}} \cdots \frac{P_{t+k} + D_{t+k}}{P_{t+k-1}} \\ &= (1 + R_t) \cdot (1 + R_{t+1}) \cdots (1 + R_{t+k}) \\ &= \prod_{i=0}^{k} (1 + R_{t-i}) \end{aligned} \tag{5.11}$$

similarly, a k-period log-return with dividends

$$\begin{aligned} r_t(k) &= log\{\frac{P_{t+1} + D_{t+1}}{P_t} \cdot \frac{P_{t+2} + D_{t+2}}{P_{t+1}} \cdots \frac{P_{t+k} + D_{t+k}}{P_{t+k-1}}\} \\ &= log(\frac{P_{t+1} + D_{t+1}}{P_t}) + log(\frac{P_{t+2} + D_{t+2}}{P_{t+1}}) + \cdots + log(\frac{P_{t+k} + D_{t+k}}{P_{t+k-1}}) \\ &= log(1 + R_t) + log(1 + R_{t+1}) + \cdots + log(1 + R_{t+k}) \\ &= r_t + r_{t+1} + r_{t+k} = \sum_{i=0}^{k} r_{t+i} \end{aligned} \tag{5.12}$$

5.5 Stationarity

The terms non-stationary and stationary form a fundamental part of time series analysis. In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series does not change over time. It does not mean that the series does not change over time, just that the way it changes does not itself change over time. Investigating time series data for the purpose of obtaining significant properties will be meaningless if the data are non-stationary or cannot be transformed to be stationary. Using non-stationary time series data in financial models produces unreliable and spurious results and leads to poor understanding and forecasting. The solution to the problem is to transform the time series data so that it becomes stationary. Capturing and examining the properties of financial time series, therefore, requires that univariate financial time series are stationary.

Definition 5.5.1. *Let $\{X_t\}$ be a time series with $E(X_t) < \infty$. The mean function of $\{X_t\}$ is.*

$$\mu_X(t) = E(X_t) \tag{5.13}$$

The covariance function of $\{Xt\}$ is

$$\gamma_X(r, s) = cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_r - \mu_X(r))] \tag{5.14}$$

for all integers r and s.

Definition 5.5.2. *Weak stationarityy. $\{X_t\}$ is weakly stationary if*
(i) $\mu_X(t)$ is independent of t, and
(ii) $\gamma_X(t + h, t)$ is independent of t for each h.

Definition 5.5.3. *Strick stationarity. $\{X_t\}$ is a strictly stationary time series if*

$$F_x(X_1, ..., X_n) = F_x(X_{1+h}, ..., X_{n+h}) \tag{5.15}$$

for all integers h and $n \geq 1$. The two random vectors have the same joint distribution function.

Strict stationarity is the most common definition of stationarity, and it is commonly referred to simply as stationarity. It is sometimes also referred to as strict-sense stationarity or strong-sense stationarity.

Chapter 6

Risk Management

Financial risks are often related to as unpredictable movements in financial variables. These unpredictable movements can be negative development in financial markets which often is the reason for losses in portfolios and/or assets. The measure of potential losses in both portfolios and assets is considered a large proportion of risk management. In order to measure these potential losses, estimates of future volatility is a necessity. In the financial world there are many approaches to estimate these volatilities, and the simplest approach is the historical standard deviation. Volatility clustering tends to follow a certain trend or pattern, large changes tends to be followed by large changes of either sign, and small changes tends to be followed by small changes. In the financial world, there are four main types of risks, namely:

1. Market risk – risk that arises due to movement in prices of financial instruments. Market risk can be classified into two types, Directional risk, which is caused due to movement in stock price and interest rates, and Non-Directional risk which is caused by volatility risks.

2. Credit risk – risk that arises when one party fails to fulfil their obligation towards their counterparties. Credit risk can be classified into two types. Sovereign Risk, which usually arises due to difficult foreign exchange policies, and Settlement Risk which arises when one party makes the payment while the other party fails to fulfil the obligation.

3. Liquidity risk – risk that arises due to inability to execute transactions. Liquidity risk can be classified into Asset Liquidity Risk and Funding Liquidity Risk. Asset Liquidity risk arises either due to insufficient buyers or insufficient sellers against sell orders and buys orders respectively. Funding liquidity risk is the risk that a bank will be unable to pay its debts when they fall due. In simple terms, it is the risk that the bank cannot meet the demand of customers wishing to withdraw their deposits.

4. Operational risk - risk that arises out of operational failures such as mismanagement or technical failures. Operational risk can be classified into Fraud Risk, which arises due to the lack of controls, and Model risk, which arises due to incorrect model application.

All of these types of risk can be measured by the widely used Value at Risk (VaR) and Expected Shortfall (ES). Value at Risk is considered as the standard risk measure by financial investors, and it's the most used risk measure. Let's denote the value of an asset at time t by X_t, and assume that the random variable X_t is observable at time t. For a given time frame Δ, the asset return over period $[\Delta_t, \Delta_{t+k}]$ can be expressed as

$$\mathcal{G}_{t+\ell} := \mathcal{G}_{[\Delta_t, \Delta_{t+ell]=(X_{t+\ell}-X_t)} \tag{6.1}$$

The loss of the asset over the same period can be expressed as

$$\mathcal{L}_{t+\ell} := \mathcal{L}_{[\Delta_t, \Delta_{t+\ell}]} = -(X_{t+\ell} - X_t) \tag{6.2}$$

$\mathcal{L}_{[\Delta_t, \Delta_{t+\ell}]}$ is observable at time Δ_{t+k}, and is the loss distribution of our investment. In risk management, we are mainly concerned with the probability of large losses, hence the probability of the upper tail of the loss distribution. The random variables X_t and X_{t+k} represents the asset value at time t and time (t+k) respectively, and $\mathcal{L}_{t+k}$ represents the loss from this given investment. The theory in this chapter is based on the following articles: Ruppert and Matteson [2015], Kenton [2019], McNeil et al. [2005] Chen [2020]. The texts also provides for further reading.

6.1 Value at Risk

Value at risk (VaR) is a statistic that measures and quantifies the level of financial risk within portfolio or position over a specific time frame. VaR is defined as the maximum amount of money expected to be lost over specific period of time, at a pre-defined confidence level.

VaR determines the potential for loss and what the corresponding probability of occurrence for that defined loss is for an entity being assessed. For a VaR measure you have to take into consideration the amount of potential loss, the probability of occurrence for the specific loss and a specific period of time Let's say a financial institution has calculated that an asset return has a 5% value at risk of 3%. This means that there is a 5% chance that the asset will decline by 3% given a specific time frame. VaR is widely used by financial institutions to measure and control the level of risk exposure, where they can apply VaR calculations to either positions or whole portfolios. However, several problems may occur when using VaR to measure the level of financial risk. For example, let's say you calculate VaR and using statistics pulled arbitrarily from a period of low volatility, this may underestimate the potential for risk events to occur and the magnitude of those events. One other example is by using the normal distribution probabilities which doesn't account for outliers or extreme events.

Definition 6.1.1. *Consider a risky asset and a fixed time horizon Δ. Denote $F_{\mathcal{L}}(\ell) = P(\mathcal{L} \leq \ell)$ and the confidence level $\alpha \in (0,1)$. Consider $\mathcal{L}$ as the loss over the holding period Δ, then for any loss distribution, VaR_α is the α^{th} upper quantile of $\mathcal{L}$. Hence,*

$$VaR_\alpha = inf\{\ell \in R : P(\mathcal{L} \leq \ell) \leq (1-\alpha)\} = inf\{\ell \in R : F_{\mathcal{L}}(\ell) \geq \alpha\} \tag{6.3}$$

is simply the maximum expected loss of a portfolio over the time horizon Δ with certain confidence level α, where R is all the real numbers

In the HSMM framework you assign the data to a specific state, done by the mhsmm package in R, based on the chosen model with combinations of emission distribution and sojourn distribution. In this case, the emission distributions are the normal distribution, the t-distribution, the skew normal distribution and the skew t-distribution, and the sojourn distributions are the shifted Poisson and the Gamma distribution. Each of these models will assign the asset returns to different states. Separately, you calculate the VaR for each of these states and then multiply it with the weight of the state. When you combine the VaR for each state you have the VaR for the whole dataset for the given time frame.

6.2 Expected Shortfall

Conditional value-at-risk (CVaR) also known as the expected shortfall (ES), is the extended risk measure of value-at-risk and quantifies the average loss, over a specified time period, just as the VaR, but beyond a given confidence level. It is a risk assessment measure that quantifies the amount of tail risk in an investment portfolio or a position.

CVaR is derived by taking a weighted average of the "extreme" losses in the tail of the distribution of possible returns, beyond the value at risk (VaR) threshold. While VaR represents a worst-case loss associated with a probability associated with that loss and a time horizon, CVaR is the expected loss if that worst-case threshold is ever crossed. CVaR measures the expected losses that occur beyond the VaR threshold. The use of CVaR as opposed to VaR leads to a more risk-free approach in terms of risk exposure. The choice between VaR and CVaR is not always clear, but volatile and engineered investments can benefit from CVaR as a check to the assumptions imposed by VaR.

Definition 6.2.1. *For any loss distribution $\mathcal{L}$, continuous or not, the expected shortfall at confidence level α is defined as*

$$ES_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_{\mathcal{L}})du, \tag{6.4}$$

where $q_u(F_{\mathcal{L}}$ is the quantile function of $F_{\mathcal{L}}$.

Hence, the expected shortfall is related to VaR_α by the equation

$$ES_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 VaR_\alpha(\mathcal{L})du, \tag{6.5}$$

which is the average of VaR over all levels of $u \geq \alpha$ and thus look further into the tail of the loss distribution.

Chapter 7

Extenstion of the mhsmm Package in R

In the analysis we have used the mhsmm package, developed by O'Connell and Højsgaard [2011], in the procedure of estimating model parameters and inference. The mhsmm package in R performs statistical inference in both hidden Markov models and hidden semi-Markov models. One of the main features of the mhsmm package is that it is designed to allow specification of custom emission distributions. If the user can provide a density function for the emission distribution and an implementation of a M-step (maximization step) for the emission distribution, the user may use their own custom emission distribution. In our case, we are going to use 4 different emission distributions not provided in the mhsmm package. The distributions are as follows: the Normal distribution, the skew-Normal distribution, the T-distribution and the skew-T distribution. The theory in this chapter and the expansion of the custom emission distribution is based on the article O'Connell and Højsgaard [2011].

Let's start by showing how a HSMM model is defined in the mhsmm package. This example is a HSMM with T-ED, Gamma SD and K=3 states.

```
J <- 3
initial <- c(1/J, 1/J, 1/J)
B_T <- list(mu = c(0.3, 0.5, 1), sigma = c(1, 1.2, 1.4), nu = c(5, 5, 5))
d_T <- list(scale = c(10, 10, 10), shape = c(1, 1, 1), type = "gamma")
model_T <- hsmmspec(initial,
                    transition = matrix(c(0, 0.5, 0.5, 0.5, 0, 0.5, 0.5, 0.5, 0),
                    byrow = T, ncol=3),
                    parms.emis = B_T,
                    sojourn = d_T,
                    dens.emis = dTF.hsmm)
hsmm_T <- hsmmfit(x, model_T, mstep = mstep.TF, maxit = 500)
```

J is the number of states, initial is the initial distribution, parms.emis = B_T is the emission distribution parameters, sojourn = d_T is the sojourn distribution parameters, transition is the transition probability matrix, dens.emis is the density function for the emission distribution and mstep is the M-step of the EM-algortihm. The hsmmspec() function specifies the starting values for the EM-algorithm and the hsmmfit() function implements the EM-algorithm. The function hsmmfit() estimates the parameters, and the updated parameters is then used in the hsmmspec() function once more. This gives a more precise result, and the EM-algorithm converges faster.

The **T emission distribution**. The dTF.hsmm is the custom density function for the emission distribution, in this case the T-distribution. The dTF.hsmm calculates the densities for the observation vector x, given state j for the model, where the model is a hmmspec() or hsmmspec() class. This class contains an emission distribution list which should have the relevant parameters for the specific emission distribution. In this case the T-distribution, where the parameters is μ, σ and ν, where ν is the degrees of freedom and govern the kurtosis of the distribution. Inside the dTF.hsmm function we have dTF function from the gamlss package which let us define the density of the T-distribution.

```
dTF.hsmm <- function(x, j, model)
  dTF(x, mu = model$parms.emission$mu[[j]],
       sigma = model$parms.emission$sigma[[j]],
       nu = model$parms.emission$nu[[j]])
```

The mstep.TF is the M-step for the emission distribution. The Function for the M-step takes two arguments, x, the vector (or dataframe) of the observed data, and wt which is a TxK matrix representing the values in eq (3.14). The return values is a list corresponding to the $emission slot in a hmmspec or hsmmspec object. In order to make the mstep.TF function work for the T emission distribution, we have to define the TF as a gamlss.family object using the function gamlss() inside the mstep.TF function.

```
mstep.TF <- function(x, wt) {
```

```
  emission <- list(mu = list(), sigma = list(), nu = list())
  for(i in 1:ncol(wt)) {
    tmp <- gamlss(x~1, family = "TF", weights = wt[, i])
    emission$mu[[i]] <- tmp$mu.fv[1]
    emission$sigma[[i]] <- tmp$sigma.fv[1]
    emission$nu[[i]] <- tmp$nu.fv[1]
  }
  emission
}
```

For the **skew-T emission distribution** we created the density function dST.hsmm. Just as for the dTF.hsmm, the dST.hsmm function calculates the densities for the observation vector x, given state j for the model, where the model is a hmmspec() or hsmmspec() class. Inside the dST.hsmm function we have the dST1 function from the gamlss package. The function dST1() defines the Skew T Type 1 density, a four parameter density, for a gamlss.family object to be used in GAMLSS fitting using the function gamlss(), with parameters μ, σ, ν and τ

```
dST.hsmm <- function(x, j, model)
  dST1(x, mu = model$parms.emission$mu[[j]],
       sigma = model$parms.emission$sigma[[j]],
       nu = model$parms.emission$nu[[j]],
       tau = model$parms.emission$tau[[j]])
```

The mstep.ST is the M-step for the skew-T emission distribution. Just as for the mstep.TF, the mstep.ST function for the M-step takes two arguments, x, the vector (or dataframe) of the observed data, and wt which is a TxK matrix representing the values in eq (3.14). In order to make the mstep.ST function work for the skew-T emission distribution, we have to define the ST1 as a gamlss.family object using the function gamlss() inside the mstep.SN function.

```
mstep.ST <- function(x, wt) {
  emission <- list(mu = list(), sigma = list(), nu = list(), tau = list())
  for(i in 1:ncol(wt)) {
    tmp <- gamlss(x~1, family = "ST1", weights = wt[, i])
    emission$mu[[i]] <- tmp$mu.fv[1]
    emission$sigma[[i]] <- tmp$sigma.fv[1]
    emission$nu[[i]] <- tmp$nu.fv[1]
    emission$tau[[i]] <- tmp$tau.fv[1]
  }
  emission
}
```

For The **Normal emission distribution** there already exist a density function for the emission distribution and a M-step function for the emission distribution that works for both hidden Markov models and hidden semi-Markov models. However, we tried to create a new density function and M-step function based on the gamlss package for the Normal distribution to see how it performed. Just as for the dTF.hsmm, the dnorm.hsmm function calculates the densities for the observation vector x, given state j for the model, where the model is a hmmspec() or hsmmspec() class. The dnorm function inside the dnorm.hsmm function is a Normal density function which let us define the density of the Normal distribution.

```
dnorm.hsmm <- function(x, j, model)
  dnorm(x, mean = model$parms.emission$mu[[j]],
        sd = model$parms.emission$sigma[[j]])
```

The mstep.NORM is the M-step for the Normal emission distribution. Just as for the mstep.TF, the mstep.NORM function for the M-step takes two arguments, x, the vector (or dataframe) of the observed data, and wt which is a TxK matrix representing the values in eq (3.14).

```
mstep.NORM <- function(x, wt) {
  emission <- list(mu = list(), sigma = list())
  for(i in 1:ncol(wt)) {
    tmp <- gamlss(x~1, family = "NO", weights = wt[, i])
    emission$mu[[i]] <- tmp$mu.fv[1]
    emission$sigma[[i]] <- tmp$sigma.fv[1]
  }
  emission
}
```

The mstep.NORM performed well compared to the already existing mstep.norm function in the mhsmm package, as the EM-algorithm used less iteration to converge, while both gave the same log-likelihood value.

For the **skew-N emission distribution** we created the density function dSN.hsmm. Just as for the dTF.hsmm, the dSN.hsmm function calculates the densities for the observation vector x, given state j for the model, where the model is a hmmspec() or hsmmspec() class. Inside the dSN.hsmm function we have the

dSN1 function from the gamlss package. The function dSN1() defines the Skew Normal Type 1 density, a three parameter density, for a gamlss.family object to be used in GAMLSS fitting using the function gamlss(), with parameters μ, σ, ν.

```
dSN.hsmm <- function(x, j, model)
  dSN1(x, mu = model$parms.emission$mu[[j]],
       sigma = model$parms.emission$sigma[[j]],
       nu = model$parms.emission$nu[[j]])
```

The mstep.SN is the M-step for the skew-N emission distribution. Just as for the mstep.TF, the mstep.SN function for the M-step takes two arguments, x, the vector (or dataframe) of the observed data, and wt which is a TxK matrix representing the values in eq (3.14). In order to make the mstep.SN function work for the skew-N emission distribution, we have to define the SN1 as a gamlss.family object using the function gamlss() inside the mstep.SN function.

```
mstep.SN <- function(x, wt) {
  emission <- list(mu = list(), sigma = list(), nu= list())
  for(i in 1:ncol(wt)) {
    tmp <- gamlss(x~1, family = "SN1", weights = wt[, i])
    emission$mu[[i]] <- tmp$mu.fv[1]
    emission$sigma[[i]] <- tmp$sigma.fv[1]
    emission$nu[[i]] <- tmp$nu.fv[1]
  }
  emission
}
```

Chapter 8

Empirical Results

The empirical analysis employ a sample data in total of 2127 trading days, and consists of daily closing prices of a number of stocks. The data sets analyzed in this thesis are listed in Table 8.1, respectively.

1. SP500 or The S&P500 comp. Index is an American stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States.
2. The ESTX50 or EURO STOXX 50 is a stock index of Eurozone stocks and it is made up of fifty of the largest and most liquid stocks
3. The FTSE All-Share Index, originally known as the FTSE Actuaries All Share Index, is a capitalisation-weighted index, comprising around 600 of more than 2,000 companies traded on the London Stock Exchange.

Table 8.1: Data sets considered.

Name	Source	N	Time period
(1) SP500	Datastream	2127	01.11/2008 - 06.22/2016
(2) ESTX50	Datastream	2127	01.11/2008 - 06.22/2016
(3) FTSE	Datastream	2127	01.11/2008 - 06.22/2016

Table 8.1 lists the data and its sources, where N denotes the total number of daily observations within the data set. The last column represents the time line available for each data index.

We restrict the timeline for all assets to the period from 01.11/2008 to 06.22/2016, due to data computational reasons. The presentation of our empirical results is structured as follows; section 8.1 is descriptive statistics, section 8.2 is model selection and section 8.3 is the empirical analysis with four subsections, (8.3.1) VaR and ES calculation in the HMM and HSMM framework, (8.3.2) Component distribution analysis, (8.3.3) Sylized facts anaysis and (8.3.4) in-sample analysis.

8.1 Descriptive statistics

The analysis aim in the first section to identify statistical properties of our data, and therefore the data must be transformed into returns. The returns are defined as the change in the natural logarithm of each market's index price (see eq.(5.5)), where P_t is the index closing price from Datastream. The daily returns for the entire period for all the datasets are shown in Figure 8.3, and the descriptive statistics are shown in Table 8.2

Table 8.2 presents descriptive statistics for all three datasets. In the table we have number of observations, minimum and maximum values, 25% and 75% quantiles, median, mean, standard deviation, variance, skewness and kurtosis. We also present the 95% and 99% Value at Risk as well as 95% and 99% Expected Shortfall, also called Conditional Value at Risk. Notice that the mean and the median for all three datasets are approximately equal to zero which means that the highest probability of the next stock return will also be approximately equal to zero. The standard deviation is a measure of spread. A low *sd* tells us that the data is closely clustered around the mean, and a big *sd* tells us that the data are more spread out. In our case, the standard deviations varies from $1.295495 - 1.609765$ which and means that 68% of the observations, and future observations most likely, will fall into this interval, that is $0 - 1.609765$ (for ESTX50).

Further, it is often valuable when evaluating an investment to know whether the instrument we examine follow a normal distribution or has some skewness present. From our summary statistics in 8.2, we notice that

there is some skewness presents in all three datasets, which means that the form of the distribution deviates a little bit from a perfect bell-shaped curve. In general, if the skewness is inside the range of the interval $[0.5, 0.5]$, the distribution is approximately symmetric. If the skewness lies in between $[-1, -0.5]$ or $[0.5, 1]$, the distribution is moderately skewed, and if skewness is outside the interval $[1, 1]$, the distribution is highly skewed. Hence, skewness measures the degree of asymmetry in the return distribution, whereas positive skewness indicates that more of the returns are positive, and negative skewness indicates that more of the returns are negative. An investor should in most cases prefer a positively skewed asset to a similar asset that has a negative skewness.

Table 8.2: Descriptive statistics of SP500, ESTX50 and FTSE

	SP500	ESTX50	FTSE
Observations	2127	2127	2127
Minimum	-9.4695	-8.2078	-9.2645
Maximum	10.957	10.437	9.3842
25% Quantile	-0.4849	-0.7981	-0.5721
Meadian	0.0635	0.0006	0.0101
75% Quantile	0.6047	0.8553	0.6296
Mean	0.0180	-0.0022	-0.0008
SD	1.3795	1.6097	1.2954
Variance	1.9031	2.5913	1.6783
Skewness	-0.3037	0.0855	-0.0635
Kurtosis	9.6155	4.3563	7.4288
95% Value at Risk	-2.1061	-2.5568	-2.109
99% Value at Risk	-4.3662	-4.8034	-3.4909
95% Expected Shortfall	-3.4672	-3.7932	-3.1076
99% Expected Shortfall	-6.0059	-5.7619	-5.1742

Table 8.2 provides the summary of the descriptive statistics for SP500, ESTX50 and FTSE. Summary statistics: min and max value, median, mean, sd, variance and quantiles as well as skewness and kurtosis which represents the third and fourth moment of the return distribution. It also includes VaR and ES

As can be seen from the histograms in Figure 8.1, the data seems to be either left or right skewed. In other words, the distribution is asymmetrical, that is, the distribution's peak is off center towards either sides, and a tail stretches away from it. Hence, from our observations all daily returns are non-normally distributed, in the form of leptokurtosis. As we can see from Figure 8.1 the extent and direction of the skewness differs across the different datasets. Thus the normal distribution is not a good model choice to explain these datasets.

For all three datasets the kurtosis is in general high, that is, between 4.356331 and 9.615567. The kurtosis measures the concentration of the return in any given part of the distribution. High values of kurtosis indicate a departure from the normal distribution. We use the kurtosis definition where the normal distribution has kurtosis equal to 3, also called "excess" kurtosis. The "excess" kurtosis is computed by the "moment" method and a value of 3 is subtracted. Our observations indicates that all of our return series are more heavy-tailed. In general, however, a rational investor should prefer an asset with a low to negative excess kurtosis, as this will indicate more predictable returns. The major exception is generally a combination of high positive skewness and high excess kurtosis.

We further tested our data for normality by examining the Quantile-Quantile (QQ) plot, as shown in Figure 8.2. The QQ-plot is a scatterplot created by plotting two sets of quantiles, that is, the empirical and theoretical quantiles, against one another. If both the empirical and the theoretical quantiles come from the same distribution, we should see the points forming a line that's roughly straight and inside the confidence interval. The larger the departure from the reference line, the greater the evidence that the data comes from a population with a different distribution than that of a normal distribution. When observing Figure 8.2 we can easily see that the observations do not lie inside the reference lines, and deviates a lot from the line when approaching the end of the tails. This confirms that the tails of the empirical distributions are not well described by the normal distribution, and again, are rather heavy-tailed.

The descriptive statistics show that the SP500 are more extreme with respect to skewness and kurtosis than both ESTX50 and FTSE. For the SP500, the values for skewness and kurtosis are -0.3037768 and 12.649512, respectively. The corresponding values for the ESTX50 are 0.08550533 and 7.369415, and -0.06350778 and 10.44913 for FTSE. While the ESTX is skewed to the right, both SP500 and FTSE is skewed to the left and all three datasets exhibits a kurtosis which is fairly high. This indicates that the normal distribution is not

adequately gonna capture the probability that we will end up in the area where the losses occurs with low frequency and high severity, and therefore underestimate the likelihood and magnitude of these extreme tail losses.

By these findings I expect the skew-t, the skew normal and the t-distribution to perform well when using these in the HSMM framework as emission distributions.

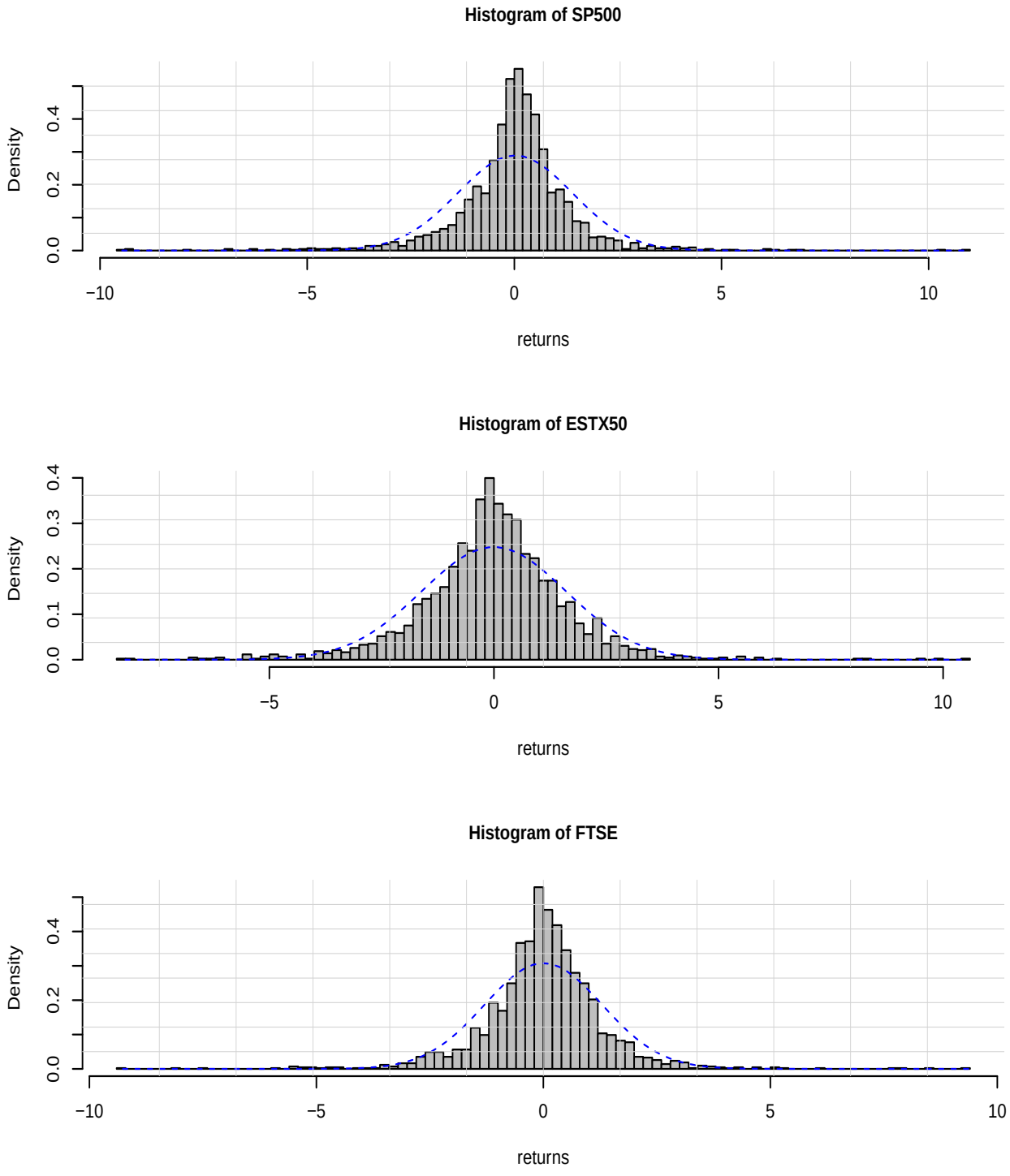


Figure 8.1: Histogram of daily compound log-returns.

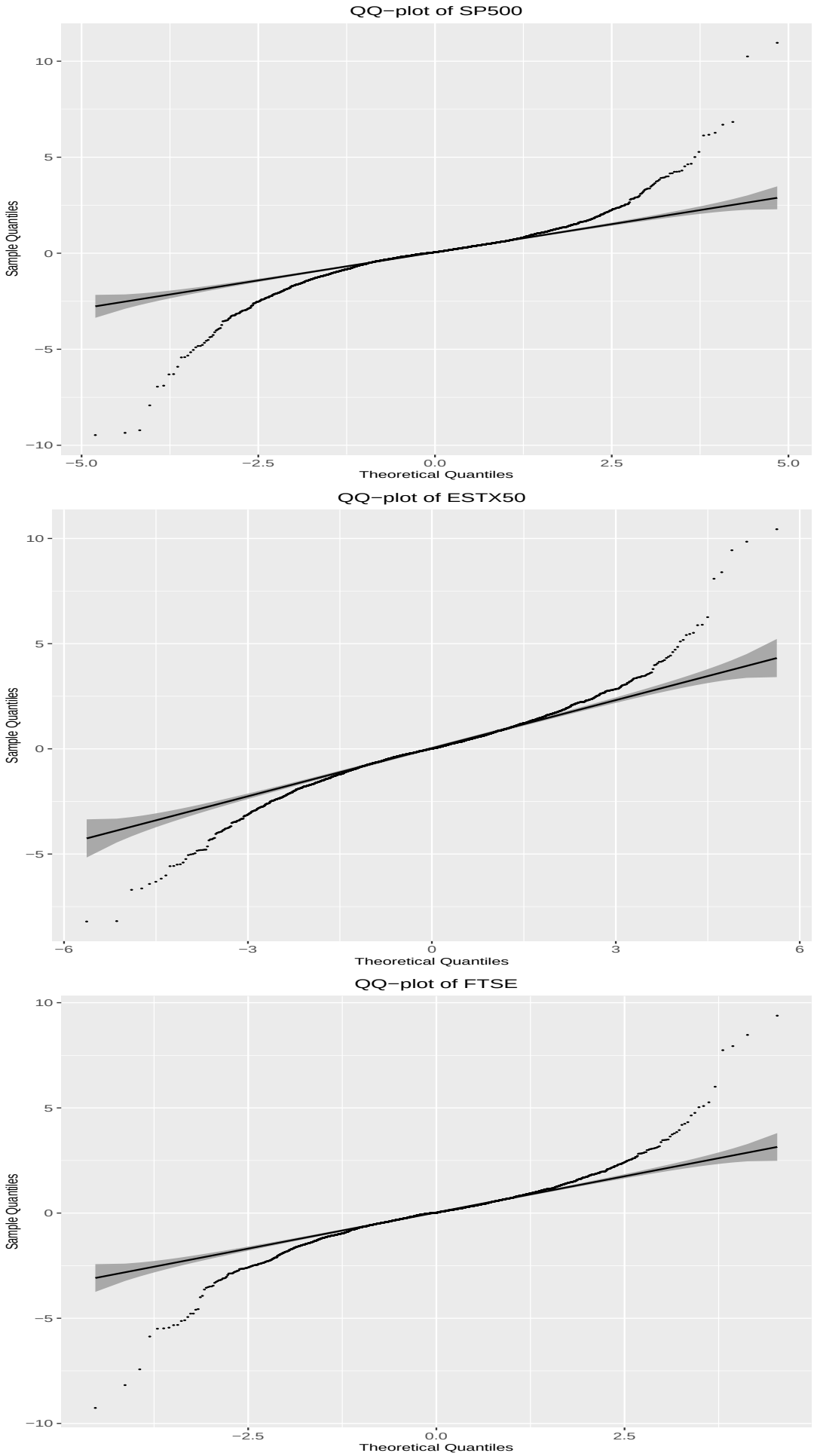


Figure 8.2: QQ-plots of daily log-return.

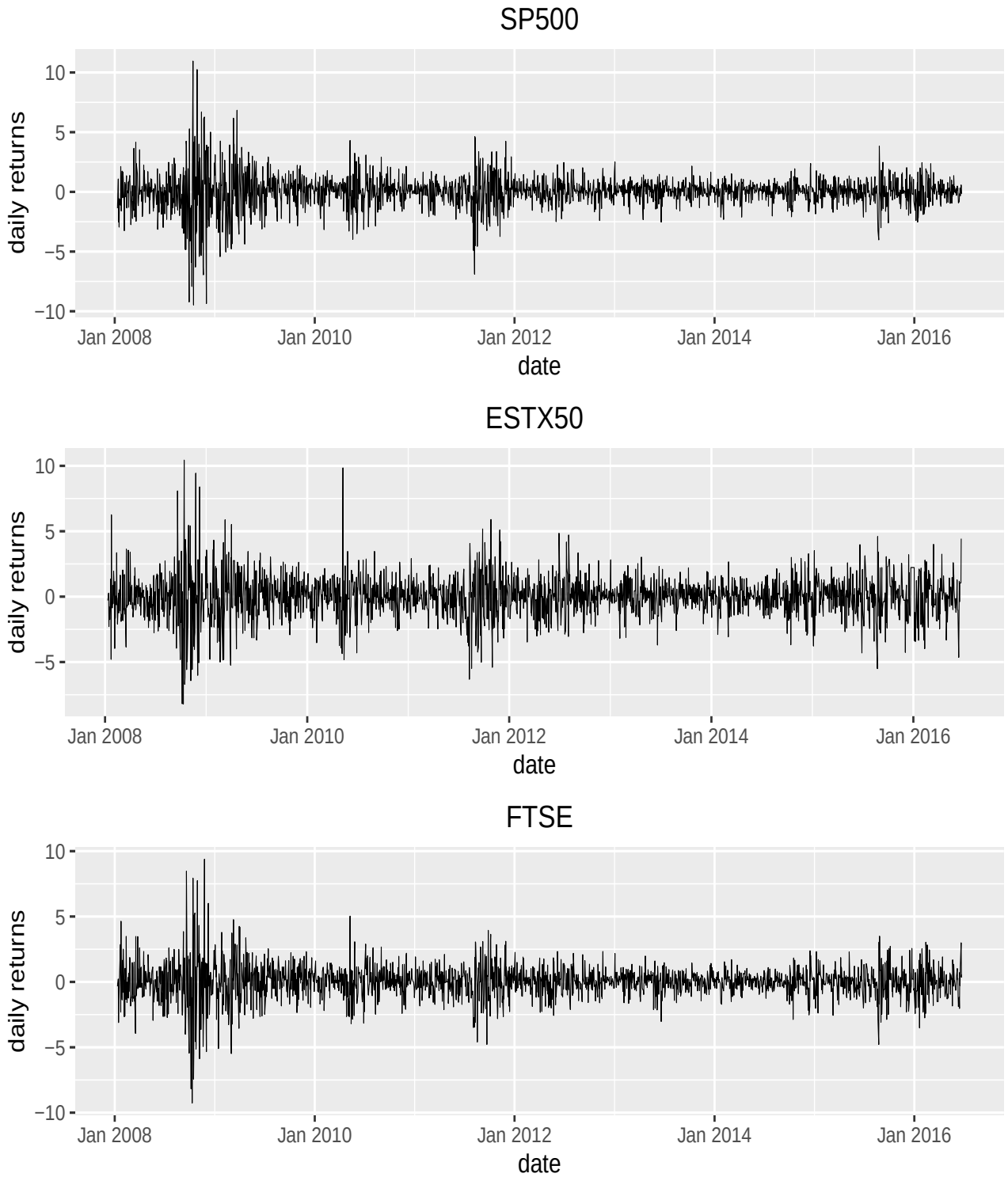


Figure 8.3: Timeseries of daily log-returns.

When analyzing financial time series, we expect the return series to display time-conditional structure of volatility, that is, periods of high or low volatility are most probably followed by high or low volatility in the following day (see Figure 8.3). Thus, strong evidence of volatility clustering supports the stylized fact that there is far more predictability in conditional volatility than in returns.

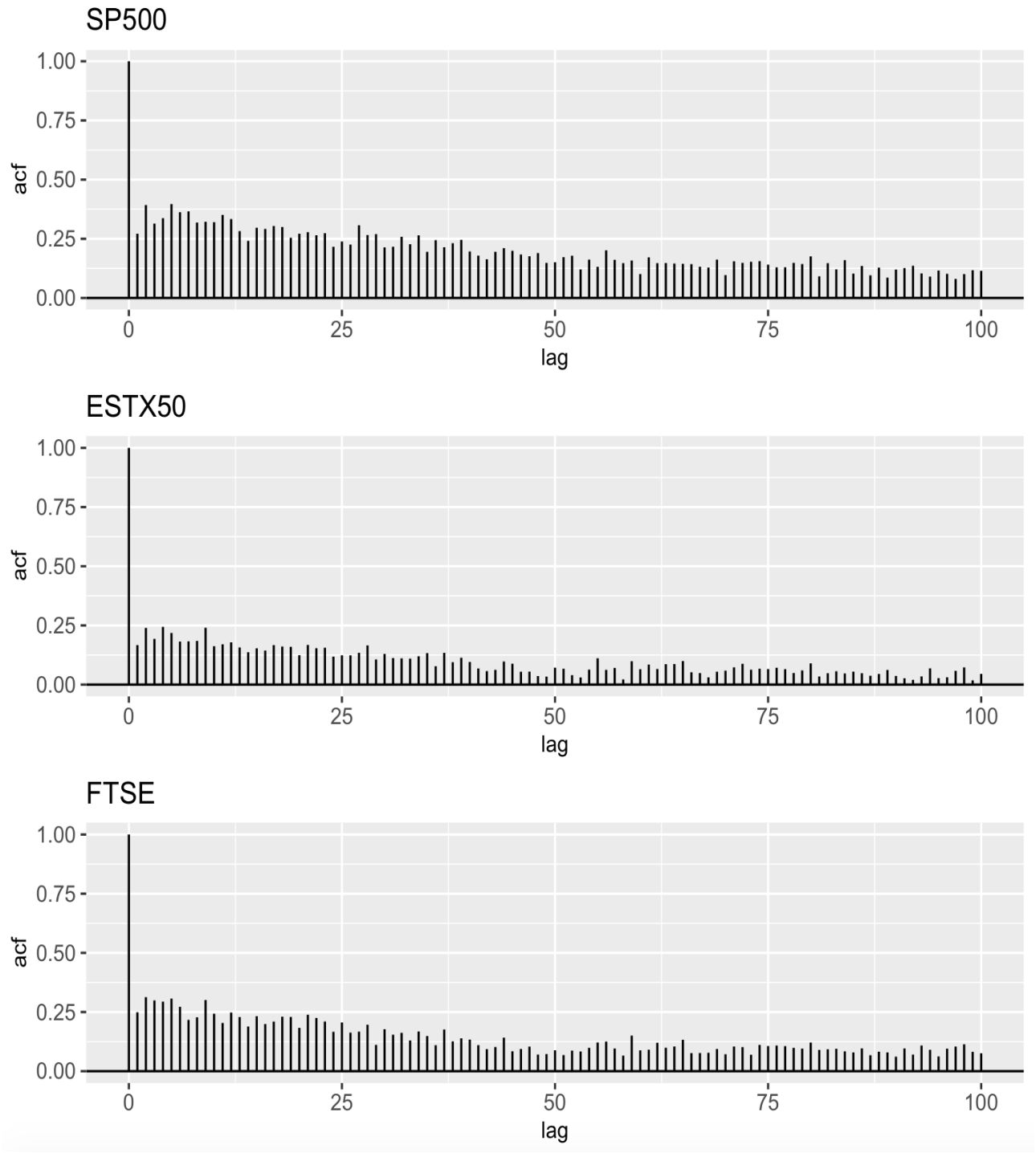


Figure 8.4: Auto-correlation plot of the absolute daily log-returns for the datasets SP500, ESTX50 and FTSE.

Another important aspect of the descriptive statistics is the stylized fact, autocorrelation, also known as serial correlation of price changes or in this case stock returns, which in general is largely insignificant, but some small and interesting anomalies do exists, as illustrated in Figure 8.4. Given that the absolute values of the stock returns can be used as a proxy for volatility, the autocorrelation of the absolute returns indicate the time dependency of the volatility. The autocorrelation portrays the correlation between observations at two different points in time, consequently, if the stock return series are independent over the given time period, the absolute values of the stock returns should not be correlated. Hence, the presence of volatility clustering in the time series can therefore be detected by a strong autocorrelation in the absolute values of the stock returns. As can be seen in Figure 8.4, there are positive values for a large number of lags. This is a good indication of volatility clustering. We will further study this stylized fact in chapter 8.3.3.

8.2 Model Selection

The theory in this chapter is based on the following articles: Costa and Angelis [2010], Goodall [2014] and Pohle et al. [2017].

Model selection procedure is a challenging topic in statistical literature and represent an essential step in hidden Markov model and hidden Semi-Markov model estimation. In the framework of HMM and HSMM, model selection is an important aspect to the choice of the number of latent states, denoted as K, of the unobserved Markov chain. The number of states should be chosen in order to enable the model to account for the dynamic pattern and the covariance structure of the observed time series. In particular, when using HMM and HSMM for exploratory purposes, the choice of K is crucial, since in many empirical developments of HMM no clue about its value is available. The consistent identification of the number of latent states, i.e. the order of the unobserved Markov chain, is a fundamental prerequisite for model parameter estimation. In order to choose the best model we will compare different information criteria on the three datasets of timeseries, SP500, ESTX50 and FTSE, respectively

Information criteria consists of two terms. A goodness of fit measure term and a penalty term. The first term, which is the goodness of fit measure term, is based on the likelihood function and is increasing by K, since adding more latent states always improves the fit of the model. The second term is the penalty term which has to be traded off against the quadratic increase in the number p of parameters that have to be estimated. The penalty is usually specified as a function of p only or as a function of both p and the number of observations T. When using the method of maximum likelihood, as the number of states in the model increases, the likelihood of the model will also increase, and the apparent fit of the model improves. However, as the number of states increases, so does the number of parameters to be estimated. For the HMM and HSMM, in addition to the emission distribution parameters, the parameters in the transition probability matrix also increase in numbers. So as the number of states increases, the numbers of parameters in the HMM's and HSMM's increase more rapidly compared to other latent variable models such as Finite mixture models.

In this paper I will compare the two most widely used information criteria, the AIC (Akaike Information criterion) and the BIC (Bayesian Information criterion).

The Akaike Information criterion is defined as:

$$AIC = -2logL + 2p \tag{8.1}$$

where L is the log-likelihood of the fitted model and p is the number of parameters. The AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. However It is also well-known that AIC is unsatisfactory because it has a tendency to overestimate the number of mixture components.
The Bayesian Information criterion is given by:

$$BIC = -2logL + p \cdot log(T) \tag{8.2}$$

with L and p as defined for the AIC and T is the number of observations. The BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model.

The BIC tends to select the simpler models (fewer parameters) than the AIC. The AIC and BIC values for the HMM's can be used to compare the fitted HMM's with different numbers of states, or models with different families of state-dependent distributions. Both criteria are based on various assumptions and asymptotic approximations. Each, despite its heuristic usefulness, has therefore been criticized as having questionable validity for real world data. But despite various subtle theoretical differences, their only difference in practice is the size of the penalty; BIC penalizes model complexity more heavily.

8.3 Empirical Analysis

In this chapter I will compare the different univariate distributions in the Hidden Markov and hidden Semi-Markov model framework to check out the stylized facts about stock return series, improve the estimates of univariate risk measures such as Value at Risk (VaR) and Expected Shortfall (ES), do an component distribution analysis and last we will do an in-sample analysis. I use 2127 daily returns for SP500, EST50 and FTSE spanning the period from January 2008 to June 2016.

As a first step of the empirical analysis, I will perform model selection by comparing HSMM's based on Normal, T, skew-N and Skew-T as emission distribution and Shifted Poisson and Gamma as the sojourn distribution. In addition I will also compare the Hidden Markov model with Normal emission distribution and geometric sojourn distribution, which is the only possible sojourn distribution possible in the HMM. Each competing model is fitted for a number of states K varying from 2 to 6, resulting in a total of 45 different models. I use the two widely-used model selection criteria to estimate the number of states K, namely AIC and BIC. The bold values in tables 8.3 - 8.8 represent the best value of considered criteria for each combination of sojourn distribution and emission distribution. Based on the results from the model selection, I will choose the three models that performed best for each of the three datasets SP500, ESTX50 and FTSE, that is, the best fitted HSMM model with shifted poisson SD, the best fitted HSMM model with Gamma SD and the best fitted HMM model to continue analysing further aspects.

Table 8.3 - 8.8 show, among the negative log-likelihood, AIC and BIC, the 95% VaR and the 95% ES for all the 45 fitted models. However, in chapter (8.3.1), I will thoroughly explain the approach used to calculate The VaR and ES and talk about the values for the best models. I will then do a component distribution analysis of one of the best 3-state HSMM models, based on model selection criteria, on the ESTX50 dataset, to highlight some important aspects and show how you can use the HSMM model to potentially improve investment in financial time series, chapter (8.3.2). I will then continue to analyse the stylized facts of stock returns in financial timeseries, chapter (8.3.3), for the three best fitted models on each dataset. And then I will finish off with an in-sample analysis, chapter (8.4.4) concerning the empirical cumulative distribution function (ECDF), to see how well the models reproduce the original data.

Table 8.3: Hidden Markov model, SP500

Normal emission distribution	2 states	3 states	4 states	5 states	6 states
Loglik	-3176.770	-3079.321	-3060.883	-3040.424	-3008.993
AIC	6371.54	6186.64	6157.76	6128.84	**6071.98**
BIC	6422.50	6265.91	6259.69	6264.74	**6224.873**
95% VaR	-1.9777	-1.8828	-1.6782	-1.5504	-1.5223
99% VaR	-3.2461	-2.8677	-2.3403	-2.2724	-2.08004
95% ES	-2.7590	-2.4694	-2.0936	-1.9930	-1.8783
99% ES	-3.8673	-3.2697	-2.7072	-2.5866	-2.274063

Table 8.3: Comparison of the Hidden Markov model with different number of states for the SP500 dataset.

Table 8.4: Hidden Markov model, ESTX50

Normal emission distribution	2 states	3 states	4 states	5 states	6 states
Loglik	-3787.45	-3745.44	-3722.55	-3712.42	-3695.57
AIC	7592.90	7518.89	7485.09	7476.85	**7453.141**
BIC	7643.86	**7598.16**	7598.34	7624.08	7628.678
95% VaR	-2.460	-2.4069	-2.2889	-2.3535	-2.197953
99% VaR	-3.6999	-3.2750	-3.1861	-3.3392	-3.0688
95% ES	-3.2333	-3.0107	-2.8821	-2.9710	-2.777328
99% ES	-4.2147	-3.8361	-3.6223	-3.8103	-3.5095

Table 8.4: Comparison of the Hidden Markov model with different number of states for the ESTX50 dataset.

Table 8.5: Hidden Markov model, FTSE

Normal emission distribution	2 states	3 states	4 states	5 states	6 states
Loglik	-3209.68	-3149.93	-3127.63	-3114.15	-3082.01
AIC	6437.37	6327.87	6291.27	6282.31	**6218.023**
BIC	6488.33	6407.14	6393.19	6435.19	**6370.91**
95% VaR	-1.9201	-1.8866	-1.7852	-1.4094	-1.682396
99% VaR	-2.9540	-2.7252	-2.5875	-2.1023	-2.2823
95% ES	-2.6175	-2.4695	-2.3616	-1.9380	-2.087761
99% ES	-3.5432	-3.1675	-3.0623	-2.5746	-2.7405

Table 8.5: Comparison of the Hidden Markov model with different number of states for the FTSE dataset.

The HMM. For The SP500 dataset in table 8.3, both the AIC and BIC selects K=6 states. The same goes for the FTSE dataset in table 8.5, both criteria selects K=6 states. For The ESTX50 dataset in table 8.4 the AIC selects K=6 states. However, it is well-known and also mentioned in Section (8.2 Model Selection), that AIC is unsatisfactory because it overestimates the number of mixture components and does not give a sufficient penalty when mixture components are high, for this reason we will not give the AIC criteria as much weight as the BIC criteria moving forward. The BIC, in this case, selcts K=3 states. The BIC for K=3 states is just slightly lower (7598.167 compared to 7598.349) than for K=4 states, while the AIC is considerably lower for K=4 states than for K=3 states. So we choose K=4 states for the ESTX50 dataset. The three models we will use further in the analysis is HMM_NO6 for SP500 and FTSE and HMM_NO4 for ESTX50.

In the HSMM framework, there are 40 combinations of models that are being fitted on the dataset. The states K are varying from 2 to 6, there are 4 types of emission distributions (ED) and two types of sojourn distributions (SD). The bold values represent the best value of considered criteria for each combination of sojourn distribution and emission distribution. Neither the AIC nor BIC chose the same number of states for any the models, so we have to argue on which model to choose based model selection criteria theory.

For the **SP500 dataset**, the shifted Poisson sojourn distribution with the Normal emission distribution gives the lowest value for the BIC, which we considered to be the most important criteria when choosing a model. The BIC selects K=5 states, which is the lowest of all the combinations of shifted Poisson SD with the four different emission distributions. The AIC selects K=6 states in this case, which is just lower than AIC value for K=5 states, however, we will proceed with the model selected by the BIC. The best fitted model for the SP500 dataset with shitfed Poisson SD is the Normal ED with K=5 states, which we will call **HSMM_NO5**.

Turning to the Gamma SD for the SP500 dataset. The gamma sojourn distribution combined with the T emission distribution for K=4 states gives the lowest value for BIC. The lowest AIC value overall for the gamma SD is with Skew-N as emission distribution for K=6 states. However, The BIC value in this case is much higher than a lot of other combination as the BIC penalize the mixture components sufficiently. As we can see from Table 8.6, the AIC almost always choose K = 5 or 6 states, as this criteria lack to sufficiently penalize a high number of mixture components/ states. The best fitted model for the SP500 dataset with Gamma SD is thus, the T ED with K=4 states, which we will call **HSMM_T4ga**.

For the **ESTX50 dataset**, the Shifted Poisson SD combined with both the T emission distribution and the Normal emission distribution gives low BIC values. The shifted Poisson SD combined with The T ED selects K=3 states for the BIC which gives the lowest BIC value of the two. The shifted Poisson with the Normal ED selects K=4 states for the BIC which is just slightly higher than the former combination, and also has a higher AIC. Both the BIC and the AIC is lower for the HSMM with shifted Poisson SD combined with T ED, so we proceed with this model which chose K=3 states, which we will call **HSMM_T3**.

For the Gamma sojourn distribution, the best combination is with the Normal emission distribution selecting K=3 states for the BIC value. However, the Gamma SD combined with the T ED selects K=3 states for the BIC, which is just barely higher than the former combination. The AIC for this combination with K=3 states is also just a bit higher than for the Gamma SD and Normal ED combination with K=3 states. So, the Gamma SD combined with Normal ED is the best model, which we will call **HSMM_NO3ga**.

Table 8.6: Hidden Semi-Markov model, SP500

	2 states	3 states	4 states	5 states	6 states
Normal emission distribution					
SD: Shifted Poisson					
Loglik	-3441.12	-3169.42	-3084.27	-3052.69	-3035.50
AIC	6904.25	6372.85	6216.54	6165.39	**6149.007**
BIC	6966.54	6469.11	6352.44	**6335.26**	6369.84
95% VaR	-1.6792	-1.9418	-1.8877	-1.8459	-1.90509
95% ES	-2.3923	-2.5978	-2.3244	-2.3280	-2.4595
SD: Gamma					
Loglik	-3175.95	-3076.93	-3070.98	-3027.30	-3022.49
AIC	6373.90	6189.87	6187.96	6118.60	**6114.98**
BIC	6436.19	**6291.79**	6318.19	6299.79	6313.17
95% VaR	-1.9725	-1.8905	-1.8805	-1.8308	-1.7834
95% ES	-2.7443	-2.4653	-2.4036	-2.3596	-2.2885
T emission distribution					
SD: Shifted Poisson					
Loglik	-3313.98	-3127.62	-3072.31	-3045.15	-3025.96
AIC	6653.96	6295.25	6204.63	6160.30	**6141.92**
BIC	6727.57	6408.50	6374.51	**6358.48**	6396.73
95% VaR	-1.7842	-1.8412	-1.9004	-1.8456	-1.7646
95% ES	-3.2475	-2.6655	-2.4788	-2.3512	-2.3100
SD: Gamma					
Loglik	-3123.00	-3060.99	-3033.26	-3020.92	-3003.79
AIC	6272.00	6163.97	6122.52	**6109.84**	6141.92
BIC	6345.61	6282.89	**6281.07**	6302.37	6344.73
95% VaR	-1.9970	-1.9009	-1.8785	-1.8580	-1.8561
95% ES	-2.8155	-2.5087	-2.4151	-2.4063	-2.4683
Skew-N emission distribution					
SD: Shifted Poisson					
Loglik	-3441.12	-3155.13	-3064.85	-3045.17	-3034.855
AIC	6908.25	6350.27	6185.70	6166.35	**6157.71**
BIC	6981.86	6463.52	**6344.25**	6381.52	6406.858
95% VaR	-1.6792	-1.7947	-1.8489	-1.9425	-1.8314
95% ES	-2.3923	-2.4598	-2.4028	-2.5409	-2.3749
SD: Gamma					
Loglik	-3175.95	-3076.93	-3040.21	-3027.20	-3008.62
AIC	6377.90	6195.87	6138.42	6126.40	**6099.25**
BIC	6451.51	6314.78	**6294.97**	6330.25	6331.42
95% VaR	-1.9725	-1.8905	-1.8597	-1.8223	-1.7884
95% ES	-2.7443	-2.4643	-2.3953	-2.3251	-2.2662
Skew-T emission distribution					
SD: Shifted Poisson					
Loglik	-3312.33	-3120.60	-3071.90	-3045.17	-3029.36
AIC	6654.66	6281.20	6205.81	**6166.35**	6173.34
BIC	6739.60	6394.45	**6381.35**	6381.52	6411.16
95% VaR	-1.8211	-1.9001	-1.9100	-1.9425	-1.8391
95% ES	-3.1672	-2.6827	-2.5193	-2.5402	-2.4652
SD: Gamma					
Loglik	-3120.23	-3060.67	-3033.31	-3023.03	-3006.34
AIC	6270.47	6169.35	6128.62	6124.06	**6106.68**
BIC	6355.41	6305.25	**6304.15**	6344.90	6372.82
95% VaR	-2.0215	-1.9031	-1.8569	-1.8246	-1.7530
95% ES	-2.8111	-2.5076	-2.3658	-2.3295	-2.2629

Table 8.6: Comparison of the Hidden Semi-Markov models with different number of states, ED and SD for the SP500 dataset.

Table 8.7: Hidden Semi-Markov model, ESTX50

	2 states	3 states	4 states	5 states	6 states
Normal emission distribution					
SD: Shifted Poisson					
Loglik	-3952.64	-3793.70	-3756.33	-3732.17	-3729.63
AIC	7927.29	7621.42	7580.66	**7526.34**	7533.26
BIC	7989.57	7717.68	**7696.56**	7701.88	7742.77
95% VaR	-2.4366	-2.4105	-2.4482	-2.4899	-2.4621
95% ES	-3.3034	-3.0449	-3.2456	-3.1728	-3.1859
SD: Gamma					
Loglik	-3787.41	-3746.01	-3730.14	-3713.29	-3685.72
AIC	7596.82	7526.035	7506.29	7484.59	**7443.45**
BIC	7659.11	**7622.29**	7636.52	7648.80	7647.30
95% VaR	-2.4648	-2.3937	-2.3273	-2.2362	-2.3282
95% ES	-3.2656	-3.0514	-2.8852	-2.7232	-3.0121
T emission distribution					
SD: Shifted Poisson					
Loglik	-3875.01	-3770.31	-3738.14	-3726.74	-3708.67
AIC	7776.02	7570.63	7538.29	7523.49	**7507.35**
BIC	7849.63	**7693.88**	7713.83	7721.67	7762.16
95% VaR	-2.3289	-2.3684	-2.5488	-2.4444	-2.3310
95% ES	-3.5670	-3.1854	-3.2554	-3.1728	-3.0054
SD: Gamma					
Loglik	-3766.44	-3735.56	-3718.05	-3713.16	-3699.66
AIC	7558.89	7531.12	7492.10	7494.32	**7487.32**
BIC	7632.50	**7624.37**	7650.64	7686.85	7736.47
95% VaR	-2.4566	-2.4249	-2.3922	-2.3341	-2.4016
95% ES	-3.2642	-3.1287	-3.1687	-2.9605	-3.1819
Skew-N emission distribution					
SD: Shifted Poisson					
Loglik	-3924.00	-3790.24	-3749.10	-3728.31	-3717.28
AIC	7874.01	7620.48	7556.20	7528.62	**7518.57**
BIC	7947.62	**7716.74**	7720.41	7732.47	7756.39
95% VaR	-2.1879	-2.3966	-2.5069	-2.4699	-2.4600
95% ES	-2.7550	-3.0198	-3.1856	-3.2558	-3.1377
SD: Gamma					
Loglik	-3787.41	-3746.01	-3727.61	-3718.71	-3702.03
AIC	7600.82	7532.03	7509.23	7509.43	**7484.06**
BIC	7674.43	**7645.28**	7662.12	7713.28	7723.70
95% VaR	-2.4648	-2.3937	-2.4501	-2.3457	-2.2848
95% ES	-3.2656	-3.0514	-3.0709	-3.0021	-2.7863
Skew-T emission distribution					
SD: Shifted Poisson					
Loglik	-3873.82	-3768.56	-3736.45	-3728.31	-3715.19
AIC	7777.64	7577.13	7583.13	7528.62	**7512.38**
BIC	7862.58	**7713.37**	7725.76	7732.56	7744.54
95% VaR	-2.2861	-2.3534	-1.9353	-2.4699	-2.4968
95% ES	-3.6223	-3.1784	-3.2481	-3.2558	-3.2268
SD: Gamma					
Loglik	-3766.38	-3749.28	-3716.62	-3705.30	-3691.52
AIC	7562.76	7544.57	7497.24	7488.61	**7481.058**
BIC	**7647.70**	7674.81	7678.44	7709.44	7758.51
95% VaR	-2.4445	-2.3938	-2.3507	-2.4164	-2.4226
95% ES	-3.2766	-3.1806	-3.0963	-3.0406	-3.0249

Table 8.7: Comparison of the Hidden Markov model with different number of states for the ESTX dataset.

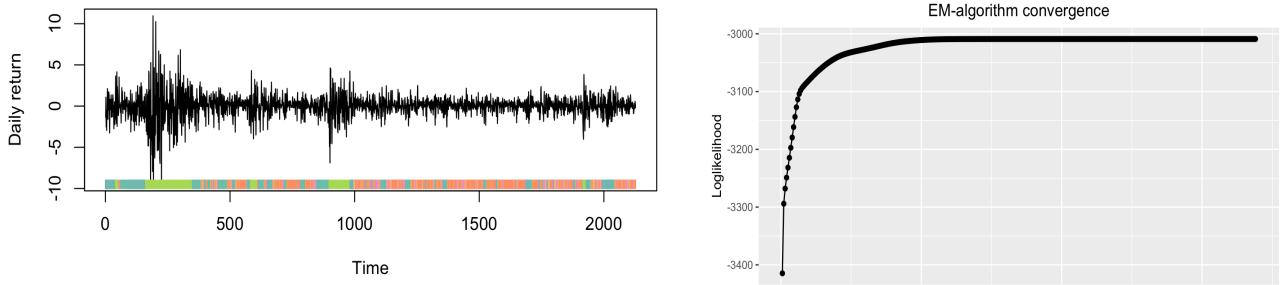
Table 8.8: Hidden Semi-Markov model, FTSE

	2 states	3 states	4 states	5 states	6 states
Normal emission distribution					
SD: Shifted Poisson					
Loglik	-3417.83	-3203.99	-3165.80	-3128.25	-3100.08
AIC	6857.67	6441.99	6381.60	6318.51	**6276.16**
BIC	6919.95	6538.25	6523.16	**6494.05**	6504.53
95% VaR	-1.8554	-1.7351	-1.8802	-1.8010	-1.8057
95% ES	-2.6279	-2.3861	-2.4150	-2.3476	-2.2004
SD: Gamma					
Loglik	-3208.59	-3150.27	-3138.54	-3093.51	-3082.85
AIC	6439.18	6334.54	6323.08	6245.01	**6235.70**
BIC	6501.47	6430.80	6453.31	**6409.23**	6433.89
95% VaR	-1.8929	-1.8543	-1.8915	-1.9042	-1.7549
95% ES	-2.5842	-2.3915	-2.3309	-2.4376	-2.1881
T emission distribution					
SD: Shifted Poisson					
Loglik	-3311.07	-3175.43	-3139.58	-3106.80	-3091.47
AIC	6648.15	6390.86	6339.16	6283.61	**6270.95**
BIC	6721.76	6504.11	6509.04	**6481.80**	6520.10
95% VaR	-1.7822	-1.7223	-1.7869	-1.7230	-1.7181
95% ES	-2.7662	-2.4152	-2.4118	-2.2621	-2.1878
SD: Gamma					
Loglik	-3174.45	-3132.63	-3106.12	-3095.11	-3102.374
AIC	6374.90	6305.27	6266.25	**6258.23**	6288.74
BIC	6448.51	**6418.52**	6419.13	6450.75	6526.57
95% VaR	-1.8979	-1.9888	-1.9070	-1.7612	-1.7678
95% ES	-2.6231	-2.5124	-2.4646	-2.2438	-2.4282
Skew-N emission distribution					
SD: Shifted Poisson					
Loglik	-3401.79	-3198.41	-3153.89	-3130.26	-3116.14
AIC	6829.59	6436.82	6361.79	6334.52	**6322.28**
BIC	6903.20	6550.07	**6514.68**	6544.03	6577.10
95% VaR	-1.5903	-1.7296	-1.9318	-1.7907	-1.8531
95% ES	-2.1595	-2.3817	-2.4212	-2.4034	-2.4213
SD: Gamma					
Loglik	-3208.59	-3150.69	-3108.65	-3092.13	-3089.44
AIC	6443.18	6343.38	6271.30	**6252.27**	6260.89
BIC	6516.79	6462.29	**6424.19**	6444.79	6493.05
95% VaR	-1.8929	-1.8551	-1.8267	-1.9117	-1.8394
95% ES	2.5842	-2.3509	-2.3062	-2.4384	-2.3217
Skew-T emission distribution					
SD: Shifted Poisson					
Loglik	-3310.98	-3172.12	-3131.15	-3130.26	-3109.15
AIC	6651.97	6390.25	6326.30	6334.52	**6298.31**
BIC	6736.91	**6497.50**	6507.50	6544.03	6524.81
95% VaR	-1.7724	-1.7746	-1.8549	-1.7907	-1.8317
95% ES	-2.7599	-2.5194	-2.4086	-2.4034	-2.4111
SD: Gamma					
Loglik	-3173.62	-3146.99	-3103.45	-3075.69	-3006.34
AIC	6377.24	6339.99	6270.91	6250.64	**6249.38**
BIC	6462.18	6470.23	**6452.10**	6471.48	6526.84
95% VaR	-1.897	-1.9693	-1.7667	-1.8465	-1.7821
95% ES	-2.6231	-2.4678	-2.2764	-2.3859	-2.2785

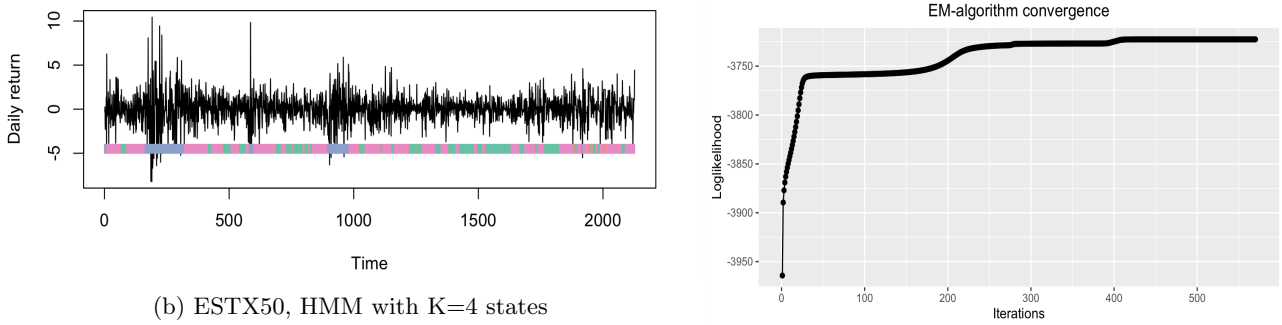
Table 8.8: Comparison of the Hidden Markov model with different number of states for the FTSE dataset.

For the **FTSE dataset**, the Shifted Poisson combined with the T ED for K=5 states gives the best value for the BIC. Two other combinations of SD and ED also worth mentioning is the Shifted Poisson SD combined with both Normal ED and Skew-T ED for K=5 and K=3 states, respectively. These models also performed well, but compared with The first combination, the AIC values of these two latter models is much higher. The model we will proceed with is **HSMM_T5**. For the Gamma sojourn distribution, the best combination is with the Normal ED selecting K=5 states for the BIC value. The best combination for the AIC is the Gamma SD and the Normal ED where the AIC selects K=6 states, but the BIC value for this model is fairly high so it's not a model we will consider. Other models also worth mentioning that performed well was Gamma SD with T ED selecting K=3 states for the BIC, and Gamma SD with Skew-N ED selecting K=4 states for the BIC. We will proceed with the model **HSMM_NO5ga**

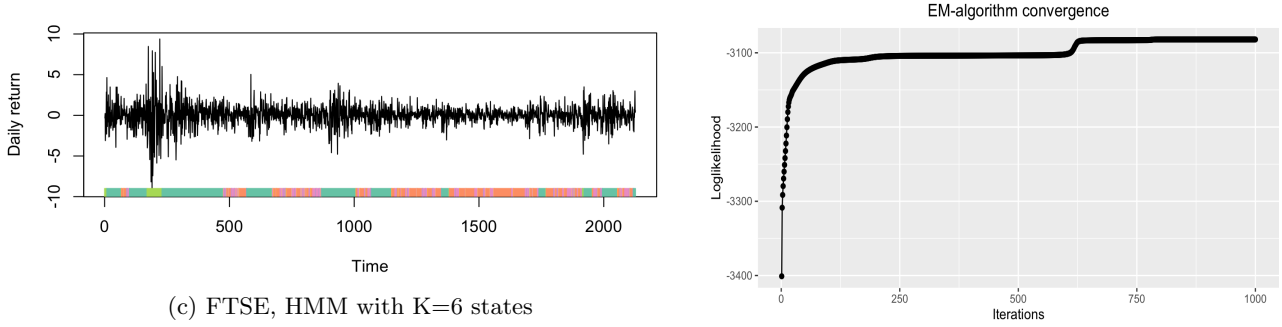
For all the models mentioned above, the predicted states combined with their EM-algorithm convergence is shown in figure 8.5, figure 8.6 and figure 8.7. As can be seen from the respective figures, the HMM models and the HSMM models with shifted Poisson sojourn distribution needed a lot more iterations for the the EM-algorithm to converge, compared to the HSMM models with Gamma sojourn distribution. The three models with Gamma SD also performed better than the three models with shifted Poisson SD when comparing the information criteria, AIC and BIC.



(a) SP500, HMM with K=6 states



(b) ESTX50, HMM with K=4 states



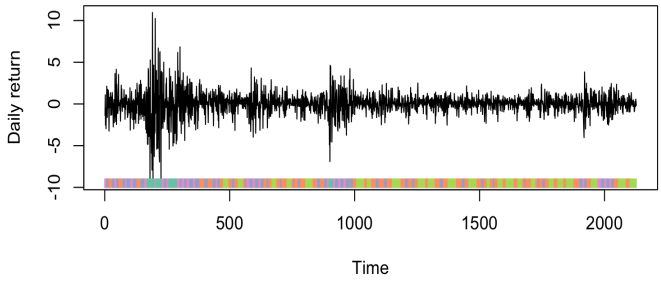
(c) FTSE, HMM with K=6 states

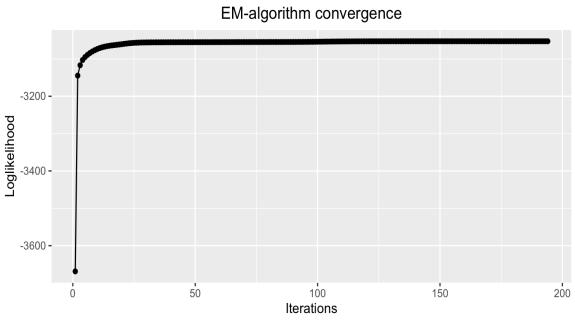
Figure 8.5: Predicted states using Viterbi-Algorithm for HMM, and the respective EM-algorithm convergence.

All three datasets either has skewness or kurtosis, or both, present. Both skew-N distribution and the T-distribution extends the Normal distribution in a way that they allow to regulate for skewness and kurtosis, respectively. The skew-T distribution has the property that it allows for regulation of both skewness and kurto-

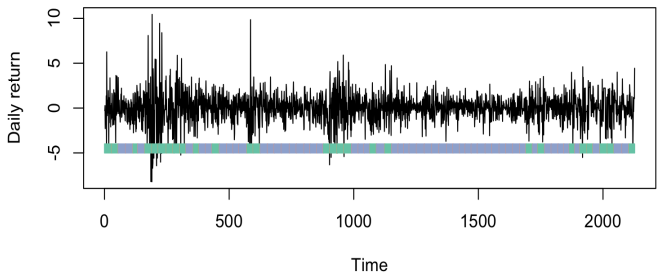
sis. As we can see from the tables 8.6, 8.7 and 8.8, concerning the HSMM models, is that the models that almost always produce the best result for the log-likelihood value (the highest value/or the lowest negative value) is the models with an emission distribution which allows for skewness, kurtosis, or both. But these models are getting a to hard penalty, that we can keep them as competing models, with a few exceptions. The reason for this is that the datasets we are analysing doesn't have enough skewness and/or kurtosis that the models with emission distribution that can regulate either of them or both performs better than the Normal-distribution, except for the T-distribution which performs just as good as the Normal. Either the skew-N distribution or the skew-T distribution as ED performed better than the Normal distribution as ED in any of the combination of models.

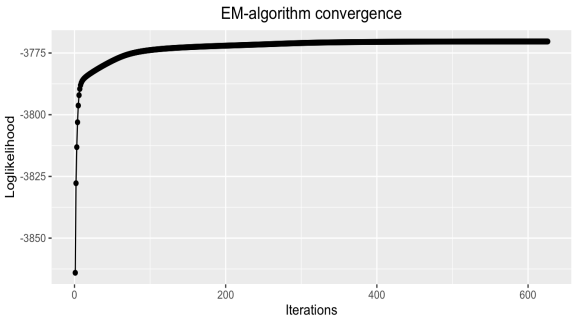
If we look at the table about the descriptive statistics, table 8.2, we can see that all the models has kurtosis present, in fact, for the SP500 dataset the kurtosis is fairly high. The skewness, however, is not that high for either of the datasets. This is the reason why the T-distribution and the Normal distribution performs better compared to the skew-N and the skew-T distribution. Even though the T-distribution is harder penalized by the AIC and BIC because of the number of parameters in the model, the kurtosis present (in some of the datasets) is large enough that it can better fit the dataset and perform better or just as good as the Normal based on the AIC and BIC. AS we can see from figure 8.6 and figure 8.7, the competing models are models with either Normal ED or T-ED.



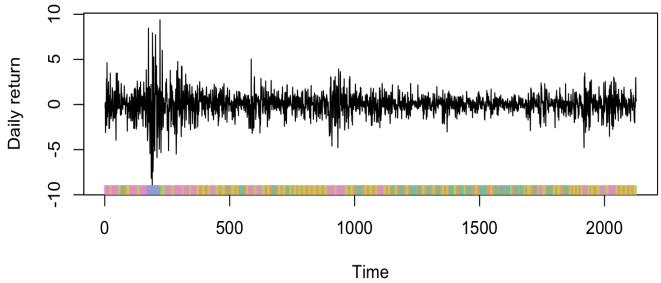


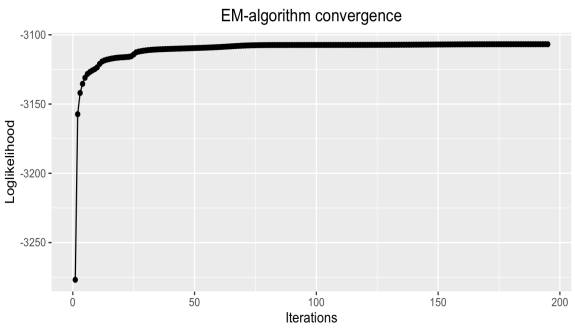
(a) SP500, HSMM with Normal ED and K=5 states





(b) ESTX50, HSMM with T ED and K=3 states





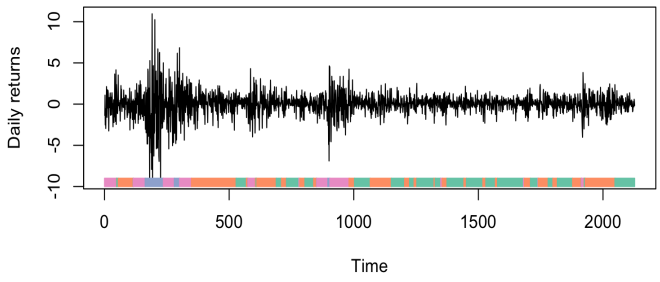
(c) FTSE, HSMM with T ED K=5 states

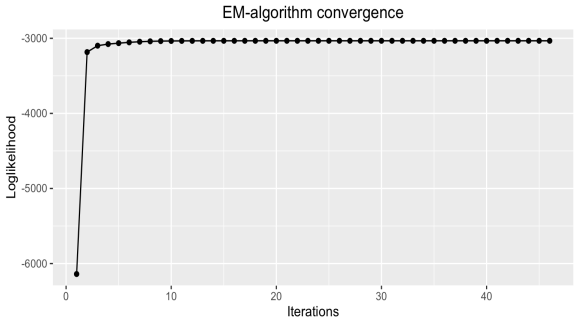
Figure 8.6: Predicted states using Viterbi-Algorithm for HSMM with Shifted Poisson sojourn distribution, and the respective EM-algorithm convergence.

When comparing the HSMM models and HMM models, you can clearly see that the HMM models most often chose models with a higher number of mixture components, in this case states. Choosing K=6 states for

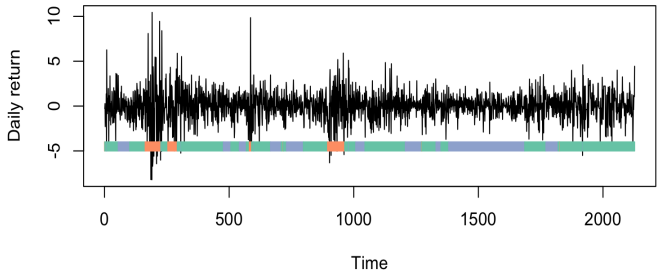
both the SP500 and FTSE dataset and K=4 states for the ESTX dataset. The reason is that when the number of states K increase, the number of parameters that are being estimated doesn't increase that rapidly for the HMM models as it does for the HSMM models. In the HMM framework, with the Normal emission distribution, you only have the sigma (σ), the mu (μ), the transition probability matrix (TPM) parameters and the initial parameters as the estimated parameters. Even though the number of states increase, the number of parameters in the model doesn't increase that much.

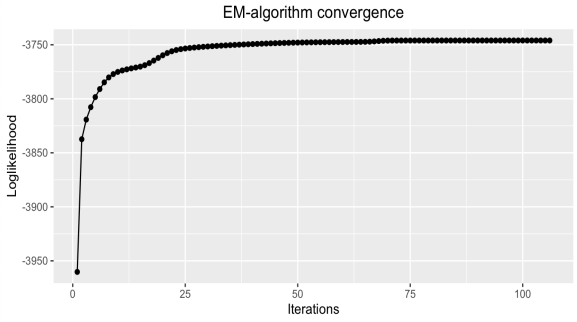
In the HSMM framework, there are two additional parameters in the model that we have to consider, namely the sojourn distribution parameters, which is the shape and the scale parameter when using the Gamma sojourn distribution, and the lambda and the shift parameter when using the shifted Poisson sojourn distribution. So when fitting, lets say a HSMM with Normal ED and Gamma SD for K=5 states, there are 10 additional parameters that are being estimated compared to the Normal ED in the HMM framework, 5 shape parameters and 5 scale parameters one for each state. This, of course, has an affect on the model selection criteria. In addition to this, we only used the Normal distribution as ED in the HMM framework, compared to the HSMM framework were we used distributions that have a lot more parameters.



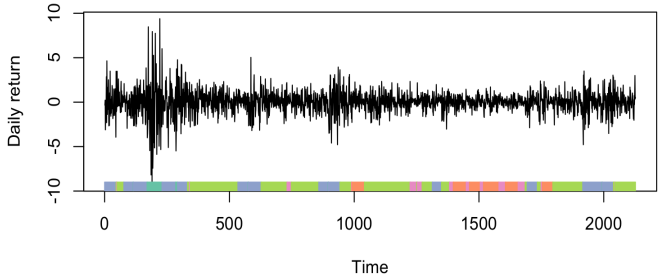


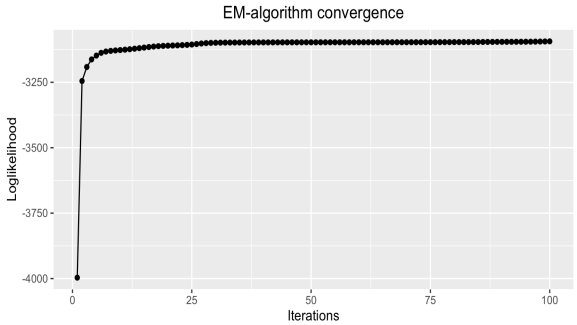
(a) SP500, HSMM with T ED and K=4 states





(b) ESTX50, HSMM with Normal ED and K=3 states





(c) FTSE, HSMM with Normal ED and K=5 states

Figure 8.7: Predicted states using Viterbi-Algorithm for HSMM with Gamma sojourn distribution, and the respective EM-algorithm convergence.

An example is the Skew-T distribution as ED, where the number of parameters consists of mu (μ), sigma (σ), a skewness-parameter (δ) and a kurtosis-parameter (κ). So when calculating the model selection criteria, AIC and BIC, they get very high. As you can see from table 8.6, table 8.7 and table 8.8, the Skew-T as emission distribution, regardless of the sojourn distribution, almost always has the highest log-likelihood. But because

of the large number of parameters that are being estimated, it almost always has the highest AIC and BIC (low values equal best fitted model), compared to the other models in the HSMM framework. Summarizing this section, the models we will use further in this thesis is:

For the SP500 dataset:

- HSMM_NO5 - shifted Poisson SD
- HSMM_T4ga - Gamma SD
- HMM_NO6.

For the ESTX50 dataset:

- HSMM_T3 - shifted Poisson SD
- HSMM_NO3ga - Gamma SD
- HMM_NO4.

For the FTSE dataset:

- HSMM_T5 - shifted Poisson SD
- HSMM_NO5ga - Gamma SD
- HMM_NO6.

8.3.1 VaR and ES calculation in the HMM and HSMM framework

The calculation of Value at Risk (VaR) in the HMM and the HSMM framework is not that complicated. We use the Viterbi-algorithm in the mhsmm package in R, created by O'Connell and Højsgaard [2011], to to find the jointly most likely configuration of states. When each of the observations is assigned to a specific state we separately calculate the Value at Risk for each of the different states. We then add up the all the VaR's for each of the states with the correct weight based on the number of observation of each state. When we combine the VaR for each state we get the VaR for the whole dataset for the given time frame. This gives a more precise result than to just calculate VaR for the whole dataset. As we can see from e.g figure 8.7 (b), it's easy to see that each state represent periods of volatility, that is, periods with high volatility or low volatility. Periods with high volatility can be divided into positive or negative frequencies of returns. The blue state represent low volatility and the green and orange state represents high volatility of frequencies of either positive or negative returns. So we have a VaR for each of these volatility periods which, combined, better capture the financial risk. Once we have calculated the value at risk, the expected shortfall (ES) is easy to obtain. While VaR represents a worst-case loss associated with a probability associated with that loss and a time horizon, ES is the expected loss if that worst-case threshold is ever crossed.

If a financial risk manager wants to know what the Value at Risk or the Expected Shortfall is for a specific asset or a portfolio of assets over a period of time in the future, the best way to do it is as follows. Firstly, fit the model on the dataset and then use the fitted model to forecast future values, based on out-of-sample forecasting (not covered in this paper). Then use the viterbi algorithm to find out the most likely configuration of states. After that, calculate the VaR for each of the periods (states) and then combine each VaR with the correct weight. Thus, this way of treating and calculating the VaR gives a more realistic result compared to the regular way when the calculation of VaR is done for the whole dataset. This is because one takes into account periods of high or low volatility. Lets say, hypothetically we "know" we are in a period of low volatility and it will last for some time, then it would not be accurate to calculate the VaR for the whole dataset since this way of calculating the VaR also includes the periods of high volatility. The most correct way would just be to calculate the VaR for that specific period. The same concept is relevant if a financial risk manager wants to know the VaR over a period of time up to a present time point.

I will now show how this calculation of VaR and ES is done in R. After the hsmmfit() function is used to implement the EM-algorithm, where the function model parameters is estimated, we use the predict() function in the mhsmm package to calculate the most likely configuration of states. This function returns a list where the component named s contains the most likely configuration of states, which is found using the Viterbi algorithm.

This arbitrary example is a HSMM with T-ED, Gamma SD and K=3 states.

```
yhat_T <- predict(hsmm_T, x)
```

Every observation is assigned to a state, so the next step of the procedure is to distinguish all the observations assigned to state 1, state 2 and state 3, and then separately calculate the Value at Risk for each specific state. The yhat_T contains a list with the states and a list with the observations, yhat_T$s and yhat_T$x, respectively. The next step is to turn these lists into a dataframe, which makes it easier for us to distinguish the states with the respective observations.

```
states <- data.frame(yhat_T$s)
obs <- data.frame(yhat_T$x)
df.states <- data.frame(states, obs)
```

```
state_1 <- df.states[df.states$yhat_T.s == 1 & df.states$yhat_T.x < 11,]
weight_1 <- length(state_1$yhat_T.x)/2127

state_2 <- df.states[df.states$yhat_T.s == 2 & df.states$yhat_T.x < 11,]
weight_2 <- length(state_2$yhat_T.x)/2127

state_3 <- df.states[df.states$yhat_T.s == 3 & df.states$yhat_T.x < 11,]
weight_3 <- length(state_3$yhat_T.x)/2127
```

Now that we have distinguished alle the observations for each state we calculate the Value at Risk for each state and then add up all the three VaR's together with the respective weights. We use the quantile function which produces sample quantiles corresponding to a given probability, in this case the 5% quantile.

```
VaR_s1 <- quantile(state_1$yhat_T.x, 0.05)
VaR_s2 <- quantile(state_2$yhat_T.x, 0.05)
VaR_s3 <- quantile(state_3$yhat_T.x, 0.05)

VaR_T <- VaR_s1 * weight_1 + VaR_s2 * weight_2 + VaR_s3 * weight3
```

We do the same for the expected shortfall, also call conditional value at risk. Once we have calculated the VaR for each state, it is easy to calculate the expected shortfall which is the expected mean loss given that a loss is occurring at or below a given quantile. This quantile is in our case the 5% Value at Risk.

```
CVaR_s1 <- mean(state_1[state_1 <= VaR_s1])
CVaR_s2 <- mean(state_2[state_2 <= VaR_s2])
CVaR_s3 <- mean(state_3[state_3 <= VaR_s3])

CVaR_T <- CVaR_s1 * weight_1 + CVaR_s2 * weight_2 + CVaR_s3 * weight_3
```

Periods of high volatility results in a higher VaR and ES value. However, they do not occur as often as more stable periods or periods of medium volatility. When we use the viterbi algorithm to compute the jointly most likely configuration of states, the observations is assigned a specific states corresponding to the degree of volatility. If we are to calculate the VaR for the whole dataset, which is done for the descriptive statistics, the periods of high volatility gets as much weight as the periods of low or medium volatility, even though the two latter periods are more frequent. On the 8.9, we have taken the VaR and ES for three best model on the three datasets combined with the ones calculated for the descriptive statistics to easily compare them.

As we can see from Table 8.9 The HMM models shows a value of VaR and ES which is much lower than that of the descriptive statistics. This can have something to do with how the computation of the states is calculated. It is well known that the sojourn time in the HMM follows a geometric distribution. This is not always desirable and limits the range of possible applications, which includes the computation of the most likely configuration of states. The HSMM models has a different setup where the sojourn times are computed explicitly and allows the models to adjust to more complex situations, which includes a better computation of the states. The HSMM models, both with Gamma SD and shifted Poisson SD, shows pretty similar results, but also shows result that are lower than for the descriptive statistics. However, these results indicates a better computation of the states. It seems like The HSMM models more accurately compute the jointly most likely configuration of states, as the VaR and ES, especially for the ESTX50 datasets, are just slightly lower than for the values calculated in the descriptive statistics. When looking at the ES which is the expected loss if that worst-case threshold i sever crossed, we can see that all the values from all models shows a result which is much lower than for the ES presented in the descriptive statistics. This shows that all models have done well dividing observations into periods of different volatility, and essentially capture the periods of high volatility, which has a great affect on the ES on the descriptive statistics. We basically give each periods of volatility the weight corresponding to the probability of their occurrence. The high-volatility periods obviously has a greater affect on the VaR and ES than the low-volatility periods, but the high-volatility periods occur less often, so they has a smaller weight on the whole computation, which gives a smoother value, especially for the ES.

Table 8.9: VaR and ES

SP500	**Des. Stat.**	**HSMM_NO5**	**HSMM_T4ga**	**HMM_NO6**
95% VaR	-2.106	-1.846	-1.879	-1.522
95% ES	-3.467	-2.328	-2.415	-1.878
ESTX50	**Des. Stat.**	**HSMM_T3**	**HSMM_NO3ga**	**HMM_NO4**
95% VaR	-2.557	-2.368	-2.394	-2.289
95% ES	-3.793	-3.186	-3.051	-2.882
FTSE	**Des. Stat.**	**HSMM_T5**	**HSMM_NO5ga**	**HMM_NO6**
95% VaR	-2.109	-1.723	-1.904	-1.682
95% ES	-3.108	-2.262	-2.438	-2.089

Table 8.9: VaR and ES for the descriptive statistics and the three best models on each of the three datasets. HSMM with Gamma SD on the second column, HSMM with shifted Poisson SD on the third column and HMM to the far right.

8.3.2 Component distribution analysis

In this analysis we are using the HSMM with Gamma SD and Normal ED for K = 3 states on the ESTX50 dataset, this is one of the the best fitted HSMM based on the model selection criteria AIC and BIC for the ESTX50 dataset. The theory and analysis of this chapter is based on Liu and Wang [2017], which gives a good introduction to HSMM in financial stock returns and also provides further reading.

The parameters on the HSMM models is estimated by the EM-algorithm, that is, the parameters for each component/state. That includes the initial parameter(s), the emission parameters, the transition probability matrix and the sojourn time distribution parameters. Table 8.10 presents the estimated parameters of the component distributions. For the parameter estimates of rest of the models, see Table A.1, A.2, A.3, A.5, A.6 and A.4 in Appendix A. Based on the estimated mean and standard deviation for each state, we compute the one-sample z-statistics in order to test the significance of the mean. The formula to compute the z-statistics is as follows

$$z_i = \frac{\overline{x_i}}{\sigma_i/\sqrt{n_i}}, i \in \{1, 2, 3\} \tag{8.3}$$

where $\overline{x_i}$ is the mean of state i, σ_i is the standard deviation of state i, and n_i is the sample size of state i. The one-sample z-test suggests that the mean of state 1 is insignificant from 0 at the 10% significant level, the mean of state 2 is significantly below 0 at the 1% significant level and the mean of state 3 is significantly above 0 at a 1% significant level.

Table 8.10: Component distribution parameters

	State 1	**State 2**	**State 3**
Mean	-0.0125	-0.2834	0.0957
Std. Dev.	1.2260	1.7436	0.9317
Sample Size	1247	183	697
Z-statistic	-0.3604	-2.1985***	2.7107***

Table 8.10: note: *$p < 0.1$ **$p < 0.5$ ***$p < 0.01$

The results indicate that the time-varying distribution of the returns depends on the hidden states, which can be interpreted as the market conditions. Specifically, state 1 corresponds to the sidewalk market, state 2 corresponds to the bear market, and state 3 corresponds to the bull market. We define the sidewalk, the bear and bull markets from the perspective of the distributional features.

Definition: A Sidewalk Market

1. The mean of the distribution of the daily returns conditional on a sidewalk market, in this case state 1, should be insignificantly different from 0.

2. It is expected to observe a roughly equal number of positive and negative daily returns in this market.
3. Because of the above statistical properties, the price in a sidewalk market mean-reversion pattern, which means that asset prices and historical returns eventually will revert to the long-run mean or average level of the entire dataset

Definition: A Bear Market

1. The mean of the distribution of the daily returns conditional on a bear market, in this case state 2, is significantly less than 0.
2. The frequency of the positive returns is expected to be smaller than that of the negative returns.
3. Because of the above statistical properties, the price in a bear market is generally decreasing.

Definition: A Bull Market

1. The mean of the distribution of the daily returns conditional on a bull market, in this case state 3, should be significantly larger than 0.
2. The frequency of the positive returns is expected to be larger than that of the negative returns.
3. Because of the above statistical properties, the price in a bull market is generally increasing

Table 8.11 presents the empirical frequency of the positive and negative returns for the fitted component distributions and confirms our interpretation of the three underlying states of the HSMM. As can be seen, the frequency of the positive return of state 3 is 54.4%, while the frequency of the negative return is 45.6%. There are negative returns in the bull market as well, but positive returns are more frequent. This statistical evidence empowers the price in the bull market to increase. Hence, state 3 can be regarded as a bull market according to its statistical features. Using the same logic, state 2 has a significant negative mean and corresponds to the bear market where the price shows a downward trend because the negative returns (57.9%) occur more often than the positive returns (42.1%). As for state 1, the frequency of the positive and negative returns is nearly the same, 49.4% and 50.6%, respectively. State 1 corresponds to the sidewalk market where the price displays a mean-reversion pattern.

Table 8.11: Frequency of Positive and Negative Returns

	State 1	**State 2**	**State 3**
Positive Return Freq.	49.4%	42.1%	54.4%
Negative Return Freq.	50.6%	57.9%	45.6%

Based on the estimated parameters in the component distribution, Figure 8.8 displays the histogram of the daily returns of the ESTX50 dataset and the density for the three fitted component distributions for the HSMM with Normal ED and Gamma SD for K = 3 states. By separating the empirical distribution into three component distributions, the HSMM is able to explain the leptokurtosis and fat tail effects. In this case, the over-peak yellow density-line in the middle part of the empirical distribution mainly results from the bull market, the mean is 0.0957 while the standard deviation is 0.868. This means that the daily returns of this state doesn't deviate that much from the mean which is positive. The bear market plays a vital role in the fat tails. The standard deviation of the bear market is 2.90134, which is much higher than for the other two markets, while the mean is -0.2834. Hence, by these component distribution parameters, the bear market is the most volatile market, followed by the sidewalk market. Surprisingly, the bull market is the most stable market.

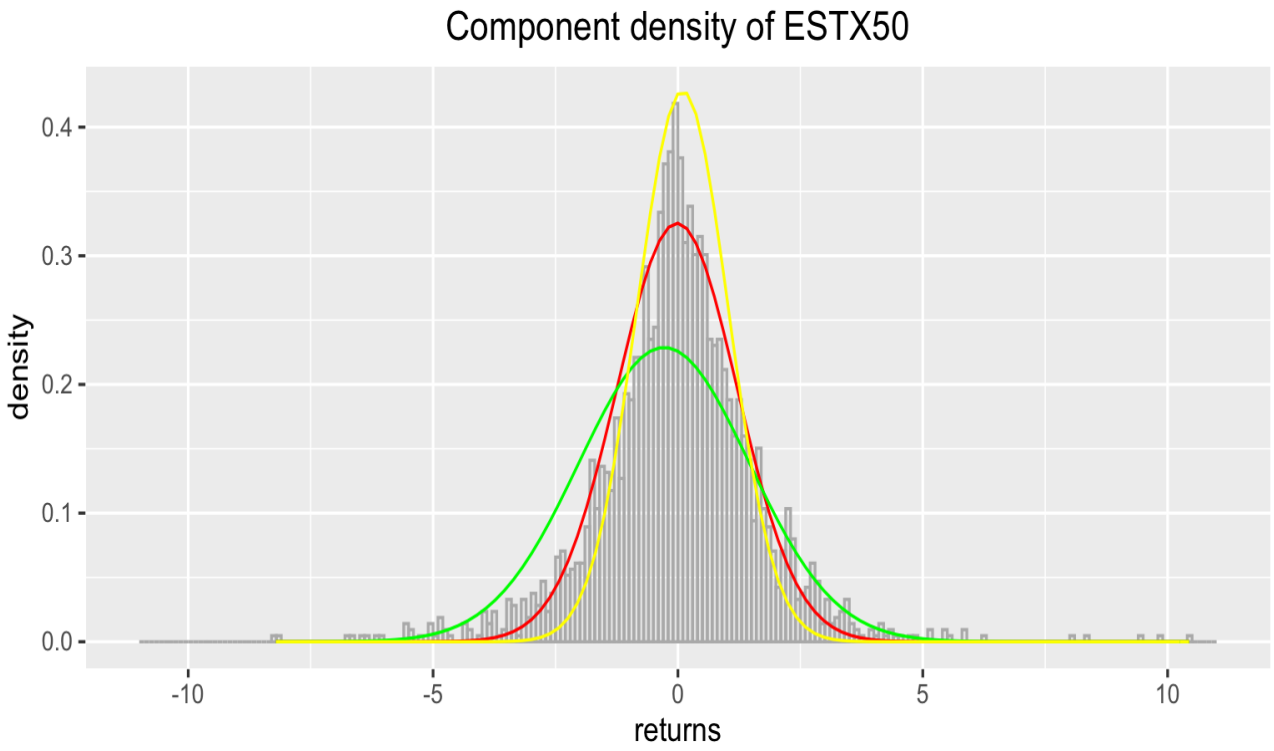


Figure 8.8: Component density for ESTX50

Figure 8.8: Shows the density-lines for the three different states of the HSMM model with Normal ED and Gamma SD. State 1 = red, State 2 = green and State 3 = yellow.

The component mean is important for price behaviour. Although the mean of state 3 (0.0957) is very small, it is still significantly larger than zero. This small but significant positive mean ensures that positive returns occur more frequently than the negative returns, which is the key feature of the bull market. The same logic can be applied to state 2, the bear market, where the mean is significantly smaller than zero and where the negative mean ensures that negative returns occur more frequently than the positive returns. The insignificant mean of state 1 ensures that its distribution is almost symmetrical around 0 and the frequency of positive returns and negative returns is nearly the same.

Table 8.12 presents the number of days, the number of times, and average sojourn time for different market conditions. Our results show that the bull market has a slightly longer sojourn time than the bear market (43.563 vs 22.875) Additionally, the average sojourn time for the sidewalk market is the longest with 56.682 days, which is much longer than in the case the other two types of markets.

Table 8.12: Sojourn time information

	State 1 (S.walk)	**State 2 (Bear)**	**State 3 (Bull)**
Number of days	1247	183	697
Number of times	22	8	16
Average sojourn	56.682	22.875	43.563

Table 8.13 gives the estimated transition probability matrix (TPM) of the HSMM for the ESTX50 dataset. The sojourn time of the HSMM is controlled by the sojourn time distribution rather than by the diagonal entries in the TPM. Hence, the diagonal entries are all zeros for the HSMM. There are a few interesting economic implications that can be drawn from the TPM. In this case, the TPM can be interpreted as follows. After a bear market, there is a 100% chanse that a sidewlak market will follow, the same goes for bull market to sidewalk market. After a sidewalk market there is a 24% chanse that a bear market will occur and a 76% chance that a bull market will occur. In other words, it is unclear, even though a bull market has highest chance of occurring after a sidewalk market, whether a bull or bear market will follow after a sidewalk market, where the price fluctuates within a certain range for a long period of time. As we can see from table 8.13 there is a zero percent chance that a bull market will occur after a bear market and vice verca. This means that after periods with either high positive return frequencies or high negative return frequencies the market must go through the

sidewalk market before proceeding to either of the two markets. This is easy to see in figure 8.7 (b).

Table 8.13: Transition probability matrix

From/to	**State 1 (S.walk)**	**State 2 (Bear)**	**State 3 (Bull)**
State 1 (S.walk)	0.00	0.24	0.76
State 2 (Bear)	1.00	0.00	0.00
State 3 (Bull)	1.00	0.00	0.00

8.3.3 Stylized facts analysis

The main focus of this third step of the empirical analysis is the stylized facts of the absolute and squared daily returns. Granger and Ding [1995a] Granger and Ding [1995b] described several temporal and distributional properties for daily return series for the SP 500 index. Rydén et al. [1998] showed that the HMM reproduces most of these properties or stylized facts. The four temporal properties are as follows:

TP1: Returns r_t are not autocorrelated (except for, possibly, at lag one).

TP2: $|r_t|$ and r_t^2 are 'long-memory', i.e., their autocorrelation functions decay slowly starting from the first autocorrelation $corr(|r_t|, |r_{t-k}|) > corr(r_t^2, r_{t-k}^2)$. The autocorrelations remain positive for many lags and the decay is much slower than the exponential rate of a typical stationary ARMA model.

TP3: The Taylor effect $corr(|r_t|, |r_{t-k}|) > corr(r_t^\theta, r_{t-k}^\theta)$, $\theta \neq 1$. Autocorrelations of powers of absolute returns are highest at power one.

TP4: The autocorrelations of sign(r_t) are negligibly small

The three distributional properties are:

DP1: $|r_t|$ and sign(r_t) are independent.

DP2: Mean $|r_t|$ = standard deviation $|r_t|$.

DP3: The marginal distribution of $|r_t|$ is exponential (after outlier correction).

Note that an exponentially distributed variable (DP3) x_t has the following properties.

PED1: $E(x_t) = Var(x_t)$ (same as DP2).

PED2: $E(x_t - E(x_t))^3 = 2$.

PED3: $E(x_t - E(x_t))^4 = 9$.

As Granger and Ding [1995a] Granger and Ding [1995b] remarked, it is not difficult to find a model possessing at least some of the above mentioned properties. From Rydén et al. [1998], elaborating Granger and Ding's example, suppose $r_t = e_t h_t$ where $\{e_t\}$ is a sequence of i.i.d. variables with zero mean independent of h_t so that TP1 holds. If the distribution of e_t is symmetric about zero then TP4 is also true.
Let $r_t = e_t h_t$ where $e_t = \text{sign}(r_t)$ and $h_t = |r_t|$. r_t, t = 1, ..., T, are drawn independently form the respective component distribution for the given models used in this paper. The HSMM with Normal emission distribution and T emission distribution with different number of states, depending on what dataset was being used, was the best fitted models. So r_t are drawn independently from one of minimum three normal or T distributions with mean assumed to be zero. By construction, TP1 holds and because the Normal's and T's are assumed to have zero means, TP4 is not violated in practice. And thus, none of our models violates TP1, TP4 or DP1.

Rydén et al. [1998] showed that a hidden Markov model with normal emission distributions and means equal to zero satisfies TP1, and that TP4 is not violated in practice. DP1 holds by construction of the model. The models fitted in our analysis allow for means unequal to zero, however, the estimation results shown in Table A.1, A.2 and A.3 show that almost all means take values very close to zero.

The mean-standard deviation ratio, skewness, kurtosis of the absolute returns estimated from the 3 datasets SP500, ESTX50 and FTSE and from the three fitted models on each dataset are presented in Table 8.14, 8.15 and 8.16. The ratio of mean and standard deviation (PED1/DP2) lies close to 1 for the original series of the three datasets. This stylized fact is reproduced very well by the HSMM models with T emission distribution. For the SP500 dataset the Mean/SD equals 1.1917, while for the HSMM_T4ga model the value is 1.2010. For the ESTX50 dataset the Mean/SD equals 1.0002, while for the HSMM_T3 model the value is 0.9899. For the FTSE dataset the Mean/SD equals 1.0860, while for the HSMM_T5 model the value is 1.0435. The HSMM models with Normal emission distribution tends to underestimate this value a bit, whereas the HMM models tend to overestimate this ratio. This confirms the analysis of Rydén et al. [1998], who noted that PED1 "has to be relaxed somewhat (the mean has to be allowed to be slightly larger than the standard deviation) if we at the same time want PED2 and PED3 to be satisfied". Furthermore, the skewness and the kurtosis are reproduced well by the HSMM models with both Normal ED and T ED. However, the HSMM models with Normal emission distribution tends to perform better, except for the FTSE dataset where the HSMM with T emission distribution produce an almost identical result, 3.0326 for original data vs 3.0360 for HSMM_T5, while the HSMM models with Normal ED perform better on the SP500 and ESTX50 dataset. The HSMM models with T-ED tend to slightly overestimate both skewness and kurtosis (except the skewness and kurtosis for HSMM_T5 on FTSE) a little bit, while the HMM models, both in regards of skewness and kurtosis, show a clear overestimation on all three datasets. Summarizing the above results it can be stated that the HSMM models reproduce PED1–PED3 comparably better than the HMM. For the SP500 dataset the HSMM_NO5 is the best model, for the ESTX50 the HSMM_NO3ga is the best model and for the FTSE dataset the HSMM_T5 is the best model.

Table 8.14: SP500

	Data	**HSMM_NO5**	**HSMM_T4ga**	**HMM_NO6**
Mean/SD	1.1917	1.1263	1.2011	2.0608
Kurtosis	17.701	19.688	20.346	22.267
Skewness	3.3274	3.4208	3.5816	4.3346

Table 8.14: Statistics of the absolute returns and the estimated models, original data mean-standard deviation ratio, skewness and kurtosis of the absolute returns estimated from the SP500 dataset and from the three fitted models HSMM_NO5, HSMM_T4ga and HMM_NO6. HSMM_NO5 = Normal ED, shited Poisson SD and 5 states. HSMM_T4ga = T ED, Gamma SD and 4 states. HMM_NO6 = Normal ED, Geometric SD and 6 states

Table 8.15: ESTX50

	Data	**HSMM_NO3ga**	**HSMM_T3**	**HMM_NO4**
Mean/SD	1.0002	0.9618	0.9899	1.5383
Kurtosis	9.7726	7.3247	13.421	16.698
Skewness	2.4123	2.2026	2.639	3.6646

Table 8.15: Statistics of the absolute returns and the estimated models, original data mean-standard deviation ratio, skewness and kurtosis of the absolute returns estimated from the ESTX50 dataset and from the three fitted models HSMM_NO3ga, HSMM_T3 and HMM_NO4. HSMM_NO3ga = Normal ED, Gamma SD and 3 states. HSMM_T3 = T ED, shifted Poisson SD and 3 states. HMM_NO4 = Normal ED, Geometric SD and 4 states

Table 8.16: FTSE

	Data	**HSMM_NO5ga**	**HSMM_T5**	**HMM_NO6**
Mean/SD	1.0860	1.0363	1.0435	2.0430
Kurtosis	15.642	18.896	18.629	55.124
Skewness	3.0326	3.1401	3.0360	6.3614

Table 8.16: Statistics of the absolute returns and the estimated models, original data mean-standard deviation ratio, skewness and kurtosis of the absolute returns estimated from the FTSE dataset and from the three fitted models HSMM_NO5ga, HSMM_T5 and HMM_NO6. HSMM_NO5ga = Normal ED, Gamma SD and 5 states. HSMM_T5 = T ED, shifted Poisson SD and 5 states. HMM_NO6 = Normal ED, Geometric SD and 6 states

Moving forward to check the stylized fact TP2, namely The slow decay of the ACF for series of abso-

lute/squared daily return. This stylized fact is difficult to model, according to Rydén et al. [1998], and stated that this stylized fact cannot be reproduced by the HMM because the decay of the autocorrelation is much faster than that observed in reality. In most cases, the ACF of the HMM model used in Rydén et al. [1998] shows a much stronger decay of the autocorrelations than the decay of the empirical ACF. Rydén et al. [1998] considered this stylized fact to be "the most difficult [...] to reproduce with a HMM". Figure 8.9, 8.10 and 8.11 shows the empirical ACF of squared and absolute returns (grey bars) as well as the ACF of the three best fitted models, according to the model selection criteria AIC and BIC. We simulate data, 20000 simulations, using simulate.hsmmspec and simulate.hmmspec in the mhsmm package, from the best three fitted models on each dataset, based on the resulted estimated parameters that arose by fitting the models on each dataset.

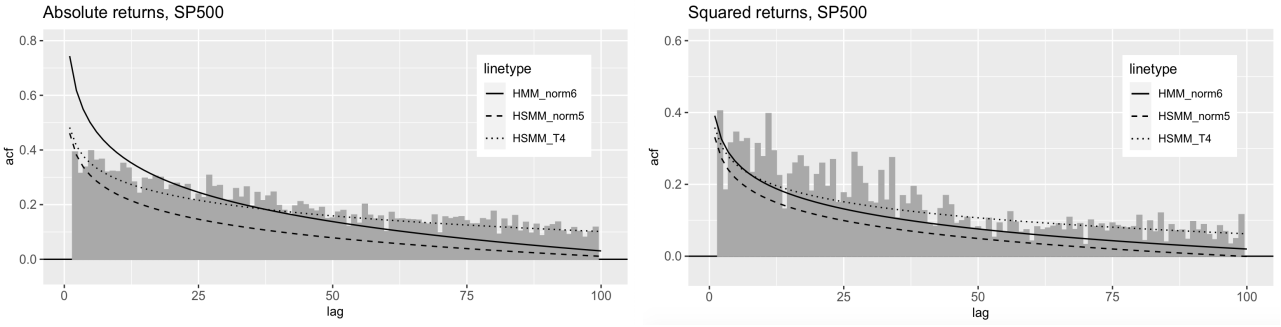


Figure 8.9: Empirical ACF and model ACF of absolute and squared returns for SP500.
Dotted line: HSMM with T ED, Gamma SD and 4 states. HSMM_T4ga = HSMM_T4
Dashed line: HSMM with Normal ED, shifted Poisson SD and 5 states. HSMM_NO5 = HSMM_norm5
Solid line: HMM with Normal ED, geometric SD and 6 states. HMM_NO6 = HMM_norm6

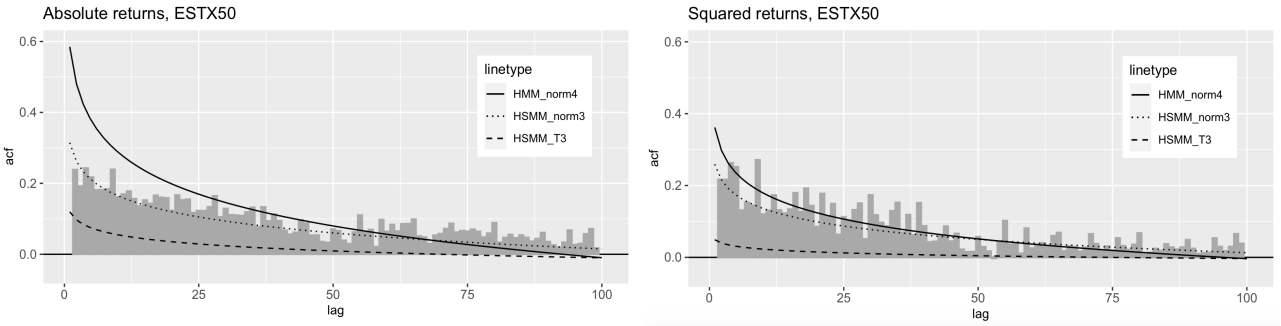


Figure 8.10: Empirical ACF and model ACF of absolute and squared returns for ESTX50.
Dotted line: HSMM with Normal ED, Gamma SD and 3 states. HSMM_NO3ga = HSMM_norm3
Dashed line: HSMM with T ED, shifted Poisson SD and 3 states. HSMM_T3
Solid line: HMM with Normal ED, geometric SD and 4 states. HMM_NO4 = HMM_norm4

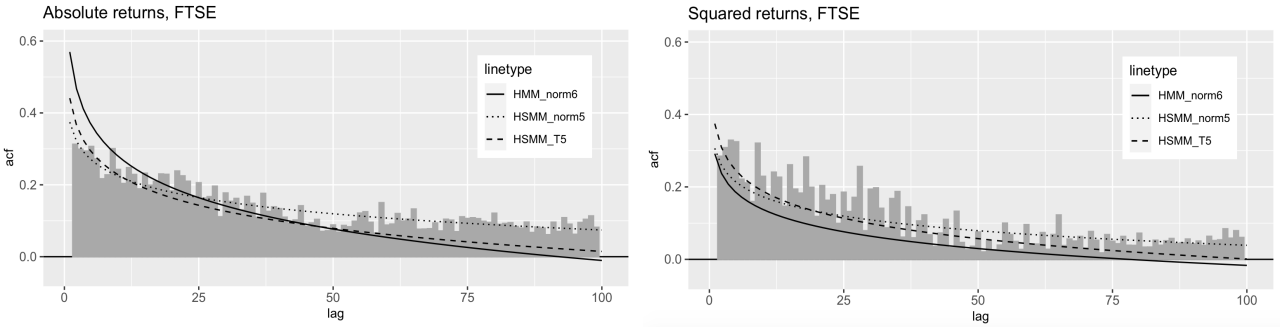


Figure 8.11: Empirical ACF and model ACF of absolute and squared returns for FTSE.
Dotted line: HSMM with Normal ED, Gamma SD and 5 states. HSMM_NO5ga = HSMM_norm5.
Dashed line: HSMM with T ED, shifted Poisson SD and 5 states. HSMM_T5
Solid line: HMM with Normal ED, geometric SD and 6 states. HMM_NO6 = HMM_norm6

The HMM models (solid line) used in our analysis are HMM's with 6,4 and 6 states for the datasets SP500, ESTX50 and FTSE respectively. The HMM's in our case doesn't show the typical strong decay of the autocorrelation which was found in Rydén et al. [1998], but is, however, more rapidly decaying than for the other two types of models. The not so strong deacying of the autocorrelation function may be caused by the number of states in the hidden Markov model framework, managing to better capture the "long memory" property of the ACF. Compared to the HMM in Rydén et al. [1998] where they used HMM models with 2-3 states, we are using 6 states for SP500 and FTSE and 4 states for ESTX50. Our results doesn't support what was found in Rydén et al. [1998], which was that the ACF of the HMM models has a strong decay. The number of states in our HMM models makes them more robust to the strong decay of the ACF, but tend to strongly overestimate the Mean-standard deviation ratio and especially the skewness and kurtosis.

For the first few lags the HMM models tend to overestimate the autocorrelation (far above the empirical ACF, grey bars) and then has a decay in the autocorrelation function, stronger than for the two other models. This is easy to see in Figure 8.9, 8.10 and 8.11 for the absolute returns. For the squared returns, the HMM models seems to perform better than they did for the absolute returns and even outperforming one of the HSMM models (HSMM_NO5), in the SP500 dataset, managing to reproduce the "long memory" property of the ACF. For the SP500 dataset, the HSMM model with T ED and Gamma SD for 4 states, namely HSMM_T4ga (dotted line) performed best and is very close to the empirical ACF, both for squared and absolute returns. The HSMM with Normal ED and Gamma SD for 5 states, namely HSMM_NO5 (dashed line) also performed very well, reproducing the "long-memory" property of the ACF, but seems to underestimate the ACF. For the ESTX50 dataset, also the HSMM model with Gamma sojourn distribution, namely the HSMM_NO3ga performed best in reproducing this stylized fact. HSMM_T3 also reproduce the "long memory" property, but looses some of its credibility due to the bad fit for the lags of lower order, and also seems to underestimate the ACF for the lags of higher order. The HMM model on absolute returns shows an overestimation for lags of low order, and then show moderate decay. For the squared returns, however, the HMM model performed better in both lags of low order and in the decay of the ACF, outperforming the HSMM_T3. For the FTSE dataset, also here, the model based on Gamma SD, in this case combined with Normal ED with 5 states (HSMM_NO5ga), performed best, showing a good fit through all the lags for both squared and absolute returns. The HSMM_T5 also performed very well, but with a slightly overestimation of the first few lags of low order for the absolute returns. The HMM model, again has an overestimation for lags of low order and shows a moderate strong decaying of the ACF.

To measure the fit of the ACF, we calculate the mean squared error (MSE) of the models and a weighted mean squared error, (wMSE). The wMSE re-weights the error at lag i by $0.95^{(100-i)}$ to increase the influence of higher order lags. The results reported in Table 8.17, 8.18 and 8.19 confirm the visual impression that the HSMM models with Gamma sojorun distribution, that is HSMM_T4ga, HSMM_NO3ga and HSMM_NO5ga, provides the best fit with respect to both criteria. The HMM models surprisingly outperforms the HSMM models with shifted Poisson sojourn distribution for the SP500 and ESTX50 dataset.

Summarizing, the six HSMM models (3 with Gamma SD and 3 with shifted Poisson SD) perform comparably better than the HMM model w.r.t distributional properties. The HSMM models with Gamma SD outperforms the HMM models and also the HSMM with shifted Poisson SD w.r.t the temporal properties, while the HSMM based on shifted poisson SD performed better in reproducing the distributional properties. Considering the fact that the HMM models performed better than the HSMM models with shifted Poisson SD for the stylized fact TP2, there wasn't a big difference in the results between the two of them. So all in all, both the HSMM's based on Gamma SD and the HSMM's based on shifted Poisson SD are outperforming the HMM models concerning the stylized facts.

Table 8.17: Mean squared error, SP500

	HSMM_NO5	**HSMM_T4ga**	**HMM_NO6**
MSE_abs x 10^3	9.3335	1.0097	7.1834
wMSE_abs x 10^3	10.532	1.0272	7.6923
MSE_sq x 10^3	7.3567	2.6353	3.0939
wMSE_sq x 10^3	5.0574	1.0433	3.6089

Table 8.17: Average mean squared error and weighted mean squared error for the ACF of squared and absolute returns for SP500 and the three fitted models.

Table 8.18: Mean squared error, ESTX50

	HSMM_NO3ga	**HSMM_T3**	**HMM_NO4**
MSE_abs x 10^3	1.0232	6.5545	5.9460
wMSE_abs x 10^3	1.4477	3.5127	1.8621
MSE_sq x 10^3	1.3194	6.9676	1.4896
wMSE_sq x 10^3	0.7750	1.9451	1.0216

Table 8.18: Average mean squared error and weighted mean squared error for the ACF of squared and absolute returns for ESTX50 and the three fitted models.

Table 8.19: Mean squared error, FTSE

	HSMM_NO5ga	**HSMM_T5**	**HMM_NO6**
MSE_abs x 10^3	0.8774	6.3274	4.9419
wMSE_abs x 10^3	0.8112	7.6817	4.7655
MSE_sq x 10^3	2.2662	3.6256	5.8646
wMSE_sq x 10^3	0.8868	3.5328	3.0173

Table 8.19: Average mean squared error and weighted mean squared error for the ACF of squared and absolute returns for FTSE and the three fitted models.

8.3.4 In-sample analysis

In-sample forecast/prediction is the process of formally evaluating the predictive capabilities of the models developed using observed data to see how effective the algorithms are in reproducing data. In this in-sample analysis I will evaluate and compare the empirical cumulative distribution function (ECDF), calculated from the original data, and compare it with the ECDF of the three best fitted models on each dataset, to see how well the data are reproduced. An empirical cumulative distribution function (ECDF) is a non-parametric estimator of the underlying cummulative distribution function (CDF) of a random variable. It assigns a probability of $1/n$ to each datum, orders the data from smallest to largest in value, and calculates the sum of the assigned probabilities up to and including each datum. The result is a step function that increases by $1/n$ at each datum. The ECDF is defined as:

$$\hat{F}_n(x) = \hat{P}_n(X \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x), \tag{8.4}$$

where $I()$ is the indicator function

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq 0 \\ 0 & \text{if } x_i > 0 \end{cases} \tag{8.5}$$

The closer the ECDF of the fitted models lie to the ECDF of the original data, the better the fitted models reproduce the data. The ECDF with green lines is the ECDF of HSMM models with shifted Poisson SD and the blue line is for the HSMM with Gamma SD. As we can see from figures 8.12, 8.13 and 8.14, the ECDF of the HSMM model, both with shifted Poisson SD and Gamma SD reproduce the original data better than that of the HMM models. To know how good each model perform in reproducing the data, we will perform a Kolmogorov-Smirnov Goodness-of-Fit Test. The D-statistic quantifies the distance between the empirical cummulative distribution function of two samples. The lower the D-statistc, the better the fitted model reproduce the original data. If the p-value is lower than the significance level, often $\alpha = 0.05$, then we reject the null hypothesis that the fitted model reproduce the original data. The alternative hypothesis is then accepted, that is, the fitted model doesn't reproduce the original data. As we can see from the tables 8.20, 8.21 and 8.22, all the fitted HMM models reject the null-hypothesis, while the HSMM models, based on either Gamma SD or shifted Poisson SD all accept the null-hypothesis. This means that all the HSMM models do well in reproducing the original data. For the SP500 dataset, HSMM_T4ga (the model with Gamma SD)

performs best, for the ESTX50 data set, HSMM_T3 (shifted Poisson SD) performs best and for the FTSE dataset, HSMM_T5 (shifted Poisson SD) performs best. Summarizing, for the in-sample analysis, the HSMM models with shifted Poisson sojourn distribution performs best in reproducing the original data.

Table 8.20: Kolmogorov-Smirnov Goodness-of-Fit Test, SP500

	HSMM_NO5	**HSMM_T4ga**	**HMM_NO6**
D-statistic	0.0207	0.0157	0.0837
p-value	0.3613	0.7202	1.69e-12

Table 8.20: Kolmogorov-Smirnov Goodness-of-Fit Test for the FTSE and the fitted models. HSMM_NO5 has shifted Poisson SD and HSMM_T4ga has Gamma SD.

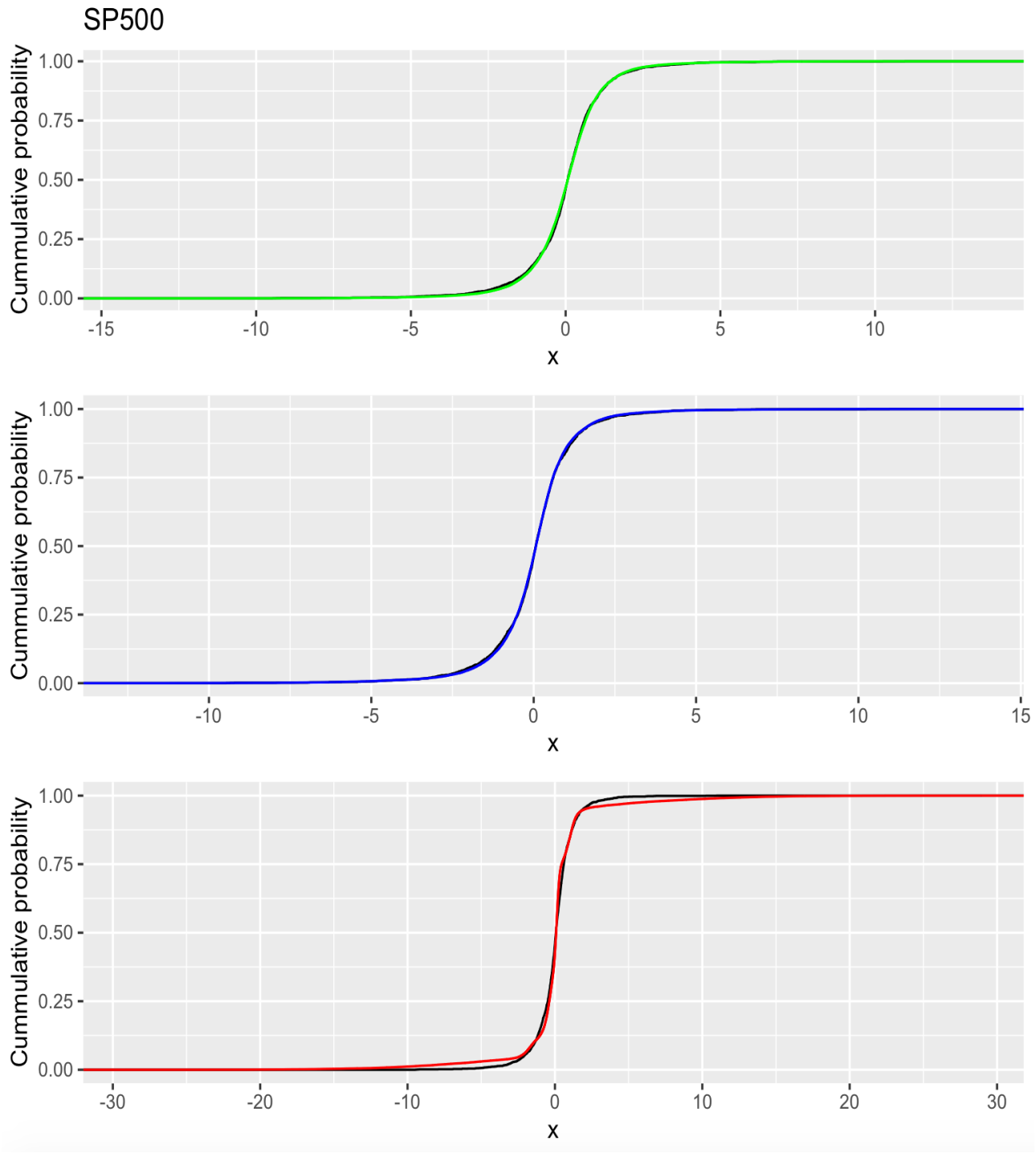


Figure 8.12: ECDF of the original data combined with the fitted ECDF of HSMM_NO5 (green line), HSMM_T4ga (blue line) and HMM_NO6 (red line).

Table 8.21: Kolmogorov-Smirnov Goodness-of-Fit Test, ESTX50

	HSMM_T3	**HSMM_NO3ga**	**HMM_NO4**
D-statistic	0.0170	0.0222	0.0681
p-value	0.6261	0.2972	3.26e-08

Table 8.21: Kolmogorov-Smirnov Goodness-of-Fit Test for the FTSE and the fitted models. HSMM_T3 has shifted Poisson SD and HSMM_NO3ga has Gamma SD.

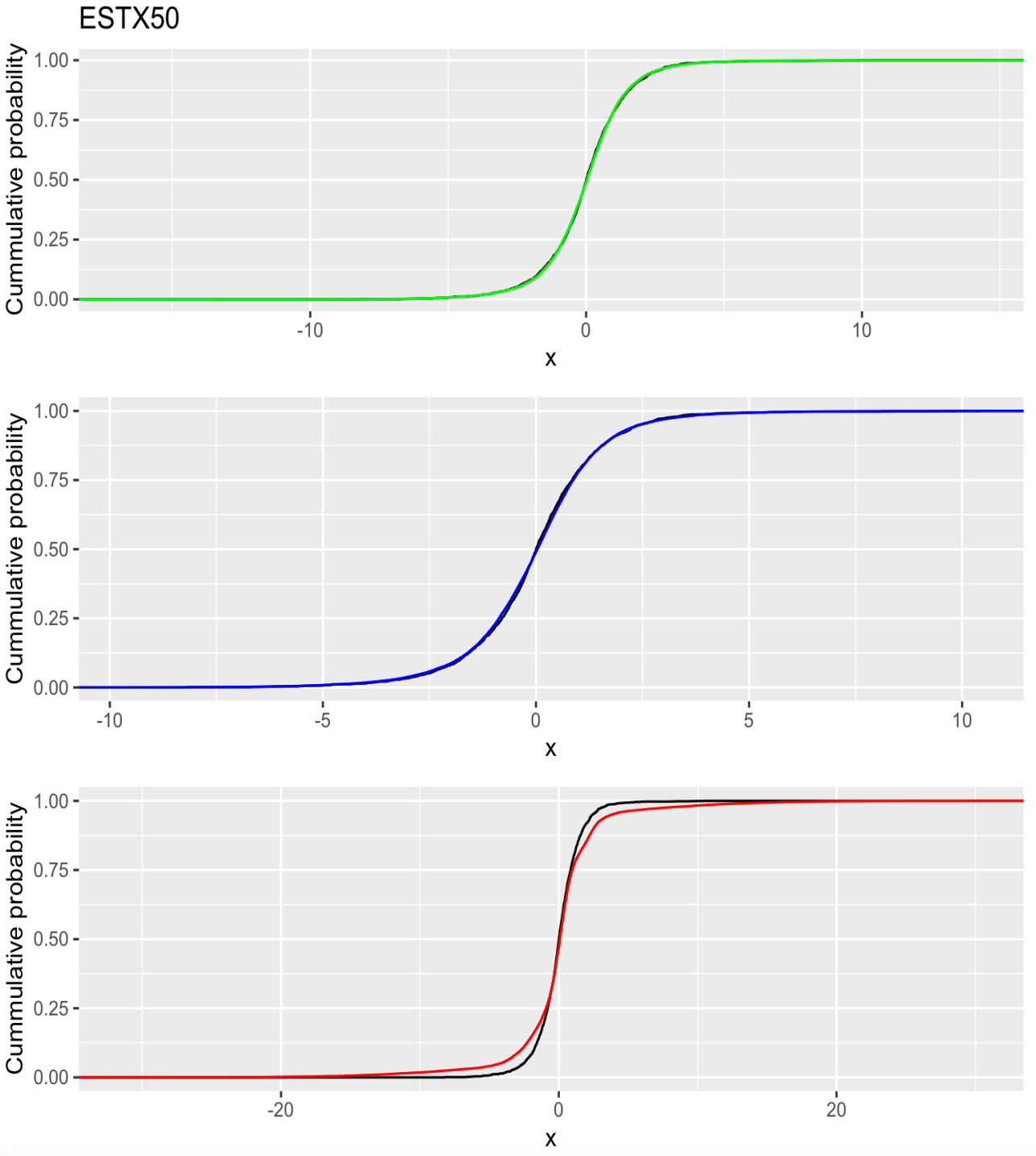


Figure 8.13: ECDF of the original data combined with the fitted ECDF of HSMM_T3 (green line), HSMM_NO3ga (blue line) and HMM_NO4 (red line).

Table 8.22: Kolmogorov-Smirnov Goodness-of-Fit Test, FTSE

	HSMM_T5	**HSMM_NO5ga**	**HMM_NO6**
D-statistic	0.0142	0.0192	0.0569
p-value	0.8205	0.4607	7.98e-06

Table 8.22: Kolmogorov-Smirnov Goodness-of-Fit Test for the FTSE and the fitted models. HSMM_NO5ga has shifted Gamma SD and HSMM_T5 has shifted Poisson SD.

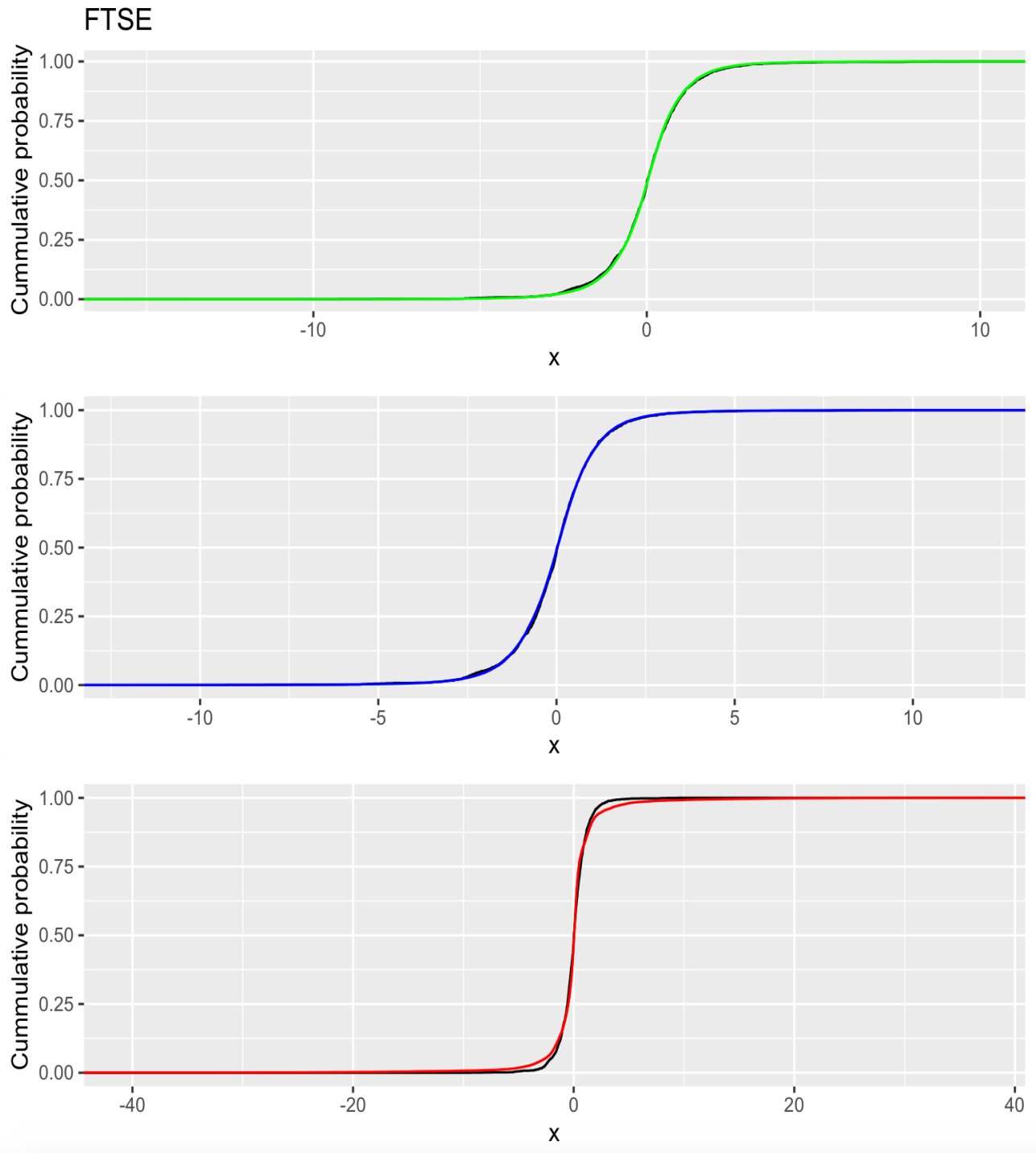


Figure 8.14: ECDF of the original data combined with the fitted ECDF of HSMM_T5 (green line), HSMM_NO5ga (blue line) and HMM_NO6 (red line).

Chapter 9

Conclusion & future work

In the past decades, the growing popularity of HMM's has led to numerous papers on applications to real-world problems. In the financial area, HMM's have regularly been employed in the context of daily returns modelling. A popular contribution to this subject was authored by Rydén et al. [1998], who showed that HMM's reproduce most of the stylized facts about daily series of returns established by Granger and Ding [1995a] and Granger and Ding [1995b], and found that the HMM couldn't reproduce the slow decay of the autocorrelation function. In Bulla and Bulla [2006], they showed that this stylized fact, which is of great importance in financial risk management, can be described better by means of HSMM's, while all other stylized facts are equally well or better reproduced. In Bulla and Bulla [2006] they used a HSMM with negative binomial sojourn time and normal conditional distributions were they also found that the HSMM reproduced the slow decay of the autocorrelation function.

In this thesis, however, we present HSMM's with a different set of combinations of sojourn distribution (SD) and emission distribution (ED), to not only analyse the stylized facts of stock return series, but to improve univariate risk measures such as Value at Risk and Expected Shortfall and to use the models to do an in-sample analysis to see how well the models reproduce the original data. In addition we have done an component distribution analysis of the HSMM with Gamma SD and Normal ED for K = 3 states on the ESTX50 dataset to show how one can interpret and identify different periods of volatilities. The sojourn distributions used in this thesis is shifted Poisson SD and Gamma SD, and combined with the four emission distributions used, Normal, skew-Normal, t, and skew-t distribution for a number of states varying from K = 2 to 6, we obtained 40 different models. In the thesis, we also included the HMM with Normal ED for K = 2 to 6 states, which resulted in 5 different models. Combined, we obtained 45 different models to test on the three datasets. The datasets used in this thesis, SP500, ESTX50 and FTSE are all analysed on the period from 01.11/2008 to 06.22/2016, which resulted in 2127 observations. To start of the analysis we applied the model selection criteria AIC and BIC to select the models we wanted to use further in our analysis. One HSMM with shifted Poisson SD, one HSMM with Gamma SD and one HMM model on each dataset. We used the widely used EM-algorithm in order to estimate the parameters of the different models.

The model selction criteria showed some interesting results. For the Gamma SD, the best fitted models was with either Normal ED or T ED for a number of states varying from K = 3 to 5, and the best fitted models with the shifted Poisson SD was also Normal ED or T ED for K = 3 to 5 states. None of the model selection criteria chose the skew-Normal or the skew-T as ED due to the large number of parameters present in these models. For the HMM the model selection criteria chose 6 states for the SP500 and FTSE dataset and 4 states for the ESTX dataset. The resulting models we further used in our analysis is as follows: for the SP500 dataset we obtained HSMM_NO5, HSMM_T4ga and HMM_NO6, for the ESTX50 dataset we obtained HSMM_NO3ga, HSMM_T3 and HMM_NO4 and for the FTSE dataset we obtained HSMM_T5, HSMM_NO5ga and HMM_NO6.

The analysis of the univariate risk measures (VaR) and (ES) showed that all the models calculated a VaR end ES, seen in Table 8.9 which was lower than that of the values calculated in the descriptive statistics, 8.2, which was calculated for the whole dataset in one go. Our approach consisted of calculating the VaR and ES for each volatility period and then combine these values with the respecting weights corresponding to each periods frequency of occurrence. This obviously reduced the VaR and ES due to the lower weights of the more volatile periods which has a great affect on these measures. We found that by separately calculating these risk measures with respect to different volatility periods, we got a more accurate result compared to the calculation done in the descriptive statistics table 8.2. In the component distribution analysis, we analysed each state distribution for the HSMM with normal ED and Gamma SD for K = 3 states, and separated them into sidewalk, bear, and bull markets. Our results indicate, supported by Liu and Wang [2017], that the time-varying distribution of

the ESTX50 stock market returns depends on the market conditions, namely the sidewalk, bear and bull market.

In the stylized fact analysis we were able to reproduce the long memory property of the stock return series, that is the slow decay of autocorrelation function. All models, seemed to perform well in doing so. However, based on visualisation from Figure 8.9, Figure 8.10 and Figure 8.11 and the values of both mean squared error and weighted mean squared erros from Table 8.17, Table 8.18 and 8.19, we found that The HSMM's based on the Gamma SD performed best and was able to very accurately produce the long memory property of the TP2 for stock returns series. Followed by the HSMM's with Gamma Poisson SD, was the HMM's and then HSMM's with shifted Poisson SD. However, the HMM's and the HSMM's with shifted Poisson produced almost similar results, with the HMM's performing just slightly better. The distributional properties of the stock returns series was well captured by the HSMM's with shifted Poisson SD, which produced the best results on the datasets SP500 and ESTX50, while the HSMM's with Gamma SD performed best on the dataset FTSE. Here, the HMM's was way behind the two other types of models, with a huge overestimation of both skewness and kurtosis and also with a mean/SD ratio that was not very well captured. The best model, when looking at the stylized facts combined, was the HSMM's with Gamma SD, followed by the HSMM's shited Poisson SD, and the the HMM's

In the in-sample analysis we tested to see how well the fitted models could reproduce the original data with comparing the EDCF of the original data with the ECDF of the fitted models. We used a Kolmogorov-Smirnov test to see how far away the ECDF of original data was from the fitted one. Here, the HSMM's with shifted Poisson SD performed best on the ESTX50 and FTSE dataset, while the HSMM's with Gamma SD performed best on the SP500 datasets. The HSMM's with both shifted Poisson SD and Gamma SD produced p-values which was way higher than the significant level (of any kind, either $\alpha = 0.1, 0.05 or 0.01$), accepting the null-hypothesis that the fitted model reproduce our original data. The HMM's produced p-values way lower than the significant level of any kind, rejecting the null hypothesis that the fitted models reproduce the original data. So the all the HSMM's performed much better than the HMM's.

In our thesis we found that the HSMM's, with either shifted Poisson SD or Gamma SD was clearly a better model choice than HMM's with Normal ED, when modelling financial timeseries. However, the easy implementation and the fast converging EM-algorithm of the HMM's makes them still an adequate model choice in regards of financial data analysis. To further study stylized facts, one can try other different combinations of sojourn distribution and emission distribution in order find a model which is superior in regards of better reproduce both the temporal and distributional properties. In our study, there wasn't a superior model, each of the HSMM's with either shifted Poisson SD or Gamma SD, and Normal ED and T ED performed good in different situations. One element that was excluded in this thesis was the out-of-sample forecast, which is an important aspect when wanting to predict future values. This can be tested out for all the models to see how well they can predict and reproduce future values. From an investment perspective, this is a very important aspect in financial modelling. The idea is to divide the dataset into a training set and a testing set. Then use the training set to fit the model up to time T, and forecast the observation at time T+1. In practice, fit the model to data from times 1:T and forecast the obs at T+1, then fit the model to data from 2:T+1 and forecast the obs at T+2, etc, using a rolling window. Then compare the forecasting values with the values on the testing set to see how well the observations are reproduced. These results can be applied to univariate risk-measures such as VaR and ES and to component distribution analysis concerning trading strategies.

Appendices

Appendix A

Estimation results

Table A.1: Parameter estimates for the HSMM models with Gamma SD

	HSMM_NO5ga	**HMM_T4ga**	**HSMM_NO3ga**
μ_1	-0.3760	0.1198	-0.0125
μ_2	-0.0251	-0.0093	-0.2833
μ_3	-0.0417	-0.4042	0.0956
μ_4	0.1371	-0.0472	
μ_5	0.0225		
σ_1^2	3.7768	0.4830	1.5031
σ_2^2	0.7134	1.0747	3.0403
σ_3^2	1.6885	3.9453	0.8681
σ_4^2	0.3769	1.8575	
σ_5^2	0.9966		
ν_1		6.3854	
ν_2		73.450	
ν_3		95.2847	
ν_4		725596278	
α_1	0.7103	5.0901	1.7654
α_2	42.234	1.0132	0.9402
α_3	46.816	1.6456	0.8285
α_4	5.2867	4.4049	
α_5	1.6205		
β_1	38.066	7.2085	20.490
β_2	1.3363	31.204	37.732
β_3	1.0305	21.061	40.909
β_4	2.7998	10.829	
β_5	39.703		

Table A.1: Estimated parameters of the HSMM with Gamma sojourn distribution. ν is the degree of freedom parameter, α and β is the shape and scale parameter, respectively, for the Gamma SD.

Table A.2: Parameter estimates for the HSMM models with shifted Poisson SD

	HSMM_NO5	**HSMM_T3**	**HSMM_T5**
μ_1	-0.3427	0.0150	0.0532
μ_2	0.0541	1.0435	0.1808
μ_3	-0.0463	-0.1626	-0.3492
μ_4	-0.1261		-0.0348
μ_5	0.1181		-0.1506
σ_1^2	3.8669	0.9412	0.4199
σ_2^2	0.9294	0.9826	0.7336
σ_3^2	1.3492	1.8971	3.8835
σ_4^2	1.9038		1.8537
σ_5^2	0.5335		1.0605
ν_1		7.9721	6.4631
ν_2		142194.2	8.9332
ν_3		5.9284	45.731
ν_4			3618053
ν_5			86804.91
λ_1	24.046	23.920	15.362
λ_2	13.551	0.0038	7.1613
λ_3	9.4550	24.861	42.126
λ_4	7.5275		25.389
λ_5	23.799		7.7098
θ_1	1	1	1
θ_2	1	1	1
θ_3	1	1	19
θ_4	1		1
θ_5	1		2

Table A.2: Estimated parameters of the HSMM with shifted Poisson sojourn distribution. ν is the degree of freedom parameter, λ and θ is the lambda and shift parameter, respectively, for the shifted Poisson SD.

Table A.3: Parameter estimates for the HMM models

	HMM_NO6	**HMM_NO4**	**HMM_NO6**
μ_1	0.3322	0.1547	0.0179
μ_2	0.1372	2.2791	0.0978
μ_3	-1.6967	-0.1804	-1.7028
μ_4	-0.2909	-0.2210	-0.1019
μ_5	-0.1223		-0.4107
μ_6	1.0586		1.4050
σ_1^2	0.9082	0.5372	2.5625
σ_2^2	0.1347	0.3259	0.2102
σ_3^2	0.4868	8.8657	0.3316
σ_4^2	0.3198	1.9311	0.5877
σ_5^2	8.1003		12.414
σ_6^2	0.2735		0.3150

Table A.3: Estimated parameters of the HMM. The model to the left is estimated on the dataset SP500, the model in the middle is estimated on the dataset ESTX50 and the model on the right is estimated on the dataset FTSE

Note that the parameters for the sojourn distribution in the HMM model is geometrically distributed, and the most likely sojourn time for every state is 1. This parameter value is not shown in the R-script in the Summary function for the hmmfit() in the R-package mhsmm.

Table A.4: TPM and Initial prob for the HSMM models with Gamma SD

HSMM_T4ga (SP500)

$$p_{ij} = \begin{bmatrix} 0.00 & 0.91 & 0.00 & 0.09 \\ 0.92 & 0.00 & 0.08 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.83 & 0.17 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 0, 1 \end{bmatrix}$$

HSMM_NO5ga (FTSE)

$$p_{ij} = \begin{bmatrix} 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.06 & 0.94 & 0.00 \\ 0.31 & 0.00 & 0.00 & 0.00 & 0.69 \\ 0.00 & 0.43 & 0.00 & 0.00 & 0.57 \\ 0.00 & 0.00 & 0.34 & 0.66 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 1, 0, 0 \end{bmatrix}$$

HSMM_NO3ga (ESTX50)

$$p_{ij} = \begin{bmatrix} 0.00 & 0.24 & 0.76 \\ 1.00 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 1, 0 \end{bmatrix}$$

Table A.4: Transition probability matrix and initial probabilities for the best fitted HSMM models with Gamma SD on the three datasets SP500, ESTX50 and FTSE

Table A.5: TPM and Initial prob for the HSMM models with shifted Poisson SD

HSMM_NO5 (SP500)

$$p_{ij} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.39 & 0.00 & 0.61 \\ 0.00 & 0.55 & 0.00 & 0.45 & 0.00 \\ 0.20 & 0.00 & 0.80 & 0.00 & 0.00 \\ 0.00 & 0.96 & 0.00 & 0.04 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 0, 1, 0 \end{bmatrix}$$

HSMM_T5 (FTSE)

$$p_{ij} = \begin{bmatrix} 0.00 & 0.64 & 0.00 & 0.00 & 0.36 \\ 0.32 & 0.00 & 0.00 & 0.00 & 0.68 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.07 & 0.00 & 0.93 \\ 0.00 & 0.82 & 0.00 & 0.18 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 0, 1, 0 \end{bmatrix}$$

HSMM_T3 (ESTX50)

$$p_{ij} = \begin{bmatrix} 0.00 & 1.00 & 0.00 \\ 0.67 & 0.00 & 0.33 \\ 0.00 & 1.00 & 0.00 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 1 \end{bmatrix}$$

Table A.5: Transition probability matrix and initial probabilities for the best fitted HSMM models with shifted Poisson SD on the three datasets SP500, ESTX50 and FTSE

Table A.6: TPM and Initial prob for the HMM models

HMM_NO6 (SP500)

$$p_{ij} = \begin{bmatrix} 0.702 & 0.000 & 0.203 & 0.000 & 0.000 & 0.095 \\ 0.000 & 0.813 & 0.000 & 0.187 & 0.000 & 0.000 \\ 0.960 & 0.000 & 0.000 & 0.000 & 0.040 & 0.000 \\ 0.000 & 0.000 & 0.099 & 0.573 & 0.000 & 0.328 \\ 0.011 & 0.006 & 0.000 & 0.000 & 0.983 & 0.000 \\ 0.000 & 0.530 & 0.000 & 0.422 & 0.000 & 0.048 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 1, 0, 0, 0 \end{bmatrix}$$

HMM_NO6 (FTSE)

$$p_{ij} = \begin{bmatrix} 0.974 & 0.000 & 0.000 & 0.000 & 0.005 & 0.021 \\ 0.000 & 0.879 & 0.000 & 0.121 & 0.000 & 0.000 \\ 0.081 & 0.000 & 0.000 & 0.887 & 0.032 & 0.000 \\ 0.000 & 0.000 & 0.101 & 0.743 & 0.000 & 0.156 \\ 0.060 & 0.000 & 0.000 & 0.000 & 0.940 & 0.000 \\ 0.000 & 0.520 & 0.000 & 0.455 & 0.000 & 0.025 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 0, 0, 1, 0 \end{bmatrix}$$

HMM_NO4 (ESTX50)

$$p_{ij} = \begin{bmatrix} 0.925 & 0.000 & 0.000 & 0.075 \\ 0.657 & 0.271 & 0.000 & 0.072 \\ 0.000 & 0.001 & 0.970 & 0.029 \\ 0.000 & 0.053 & 0.008 & 0.939 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0, 0, 1, 0 \end{bmatrix}$$

Table A.6: Transition probability matrix and initial probabilities for the best fitted HMM models on the three datasets SP500, ESTX50 and FTSE

Appendix B

Re-estimation formulae

The theory in this chapter is based on O'Connell and Højsgaard [2011], Bulla [2006] and Bulla and Bulla [2006]

B.1 State occupancy distribution

B.1.1 Shifted Poisson

The state occupancy distribution, also called the sojourn distribution, used for the hidden semi-Markov models in this paper is the shifted Poisson distribution and the Gamma distribution. These two distribution is the only available in O'Connell and Højsgaard [2011] who described the R package mhsmm The re-estimation of the parameters is done in the M-step of the EM-algorithm, not only the observation components, but also the state duration density. Guédon [2003] provides derivations for $d_i(u)$ as a non-parametric probability mass function using eq (3.13) as

$$d_i(u) = \frac{\eta_{iu}}{\sum_v \eta_{iv}} \tag{B.1}$$

One of the possibilities in the mhsmm package is to use common discrete distributions with an additional shift parameter d that sets the minimum sojourn time (d≥1), in our case, the Poisson distribution with density

$$d_j(u) = \frac{e^{\lambda}\lambda^{(u-d)}}{(u-d)!} \tag{B.2}$$

we estimate $\bar{\lambda}_i = \sum_{v=1}^{T}(v-d)\eta_{iv}$ for all possible shift parameters d = 1, ...,min$(u : \eta_{iu} > 0)$, choosing the d which gives the maximum likelihood.

B.1.2 Gamma

Another possibility in the mhsmm package in R, described by O'Connell and Højsgaard [2011], is to let the sojourn times follow a Gamma distribution, that is $U_r|S_r = i \sim \Gamma(a_i, b_i)$. The parameters are estimated as follows: The likelihood for the Gamma distribution can be maximized with respect to its parameters by solving

$$log(\hat{a}_i) - \psi(\hat{a}_i) = log(\bar{u}_i) - \overline{log(u_i)} \tag{B.3}$$

where $\psi()$ is the digamma function and the scale parameter is estimated by $\hat{b}_i = \bar{u}_i/\hat{a}_i$. We use

$$\bar{u}_i = \frac{\sum_u \eta_{iu} u}{\sum_u \eta_{iu}} \tag{B.4}$$

and

$$\overline{log(u_i)} = \frac{\sum_u \eta_{iu} log(u)}{\sum_u \eta_{iu}} \tag{B.5}$$

and then solve the equation using a numerical maximization algorithm, e.g Newton's method.

B.2 The observation component

Based on the model selection criteria AIC and BIC, we tested all of our models. That is, hidden semi-Markov models with Two different sojourn times, shifted Poisson and Gamma distribution, and four different emission distributions, Normal, skew-Normal, T and Skew-T distribution, and number states varying from 2 to 6. Based on the criteria AIC and BIC, the hidden semi-Markov models that performed best was the ones with either Normal distribution or t-distribution as the component/emission distribution. These were the models we further used in our analysis, that is chapter 8.3. We will in this appendix show the re-estimation formulas for both Normal and T distribution as the component/emission distribution. The theory in this chapter is based on Bulla and Bulla [2006] and Bulla [2006].

B.2.1 t component distribution

The derivation and maximization of the Q-function for the t distribution is not entirely straightforward, as some of the equations requires numerical maximization methods to obtain the ML estimates. However, the techniques presented by Peel and MClachlan [2000] for the estimation of mixtures of t-distributions can be adopted to the case of a HSMM. The density of the t distribution with location parameter $\boldsymbol{\mu}$, ν degrees of freedom and positive definite inner product matrix $\boldsymbol{\Sigma}$ is given by:

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+1}{2})|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1+\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Mahalanobis distance

$$\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

and p dimension of observations. As ν approaches infinity, f converges to the density function of a Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In the case of conditional t distribution, the observation distribution from Equation (3.14) is

$$b_j(\boldsymbol{x}_t) = \frac{\Gamma(\frac{\nu_j+1}{2})|\boldsymbol{\Sigma_j}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu_j}{2})\{1+\delta(\boldsymbol{x_t}, \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})/\nu_j\}^{\frac{1}{2}(\nu_j+p)}} \tag{B.6}$$

The re-estimation formulae for $\boldsymbol{\mu_j}$ and $\boldsymbol{\Sigma_j}$ can be derived explicitly, yielding

$$\boldsymbol{\mu}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)u_{jt}^{(k)}\boldsymbol{x_t}}{\sum_{t=0}^{\tau-1} L_j(t)u_{jt}^{(k)}} \tag{B.7}$$

and

$$\boldsymbol{\Sigma}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)u_{jt}^{(k)}(\boldsymbol{x_t}-\boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x_t}-\boldsymbol{\mu}_j^{(k+1)})^T}{\sum_{t=0}^{\tau-1} L_j(t)} \tag{B.8}$$

where

$$u_{jt}^{(k)} := \frac{\nu_j^{(k)}+p}{\nu_j^{(k)}+\delta(\boldsymbol{x_t}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})} \tag{B.9}$$

In Bulla [2006] he refer to Kent et al. [1994] who says that for the case of a single component t distribution, the denominator of eq (B.3) can also be replaced by $\sum_{t=0}^{\tau-1} L_j(t)u_{jt}^{(k)}$ to increase the speed of convergence. The re-estimation of the degrees of freedom ν_j is more intricate. The estimator $\nu_j^{(k+1)}$ is the unique solution of the equation

$$-\psi\left(\frac{1}{2}\nu_j^{(k)}\right)+log\left(\frac{1}{2}\nu_j^{(k)}\right)+1+\frac{1}{\sum_{t=0}^{\tau-1} L_j(t)}\left[\sum_{t=0}^{\tau-1} L_j(t)\left(logu_{jt}^{(k)}-u_{jt}^{(k)}\right)\right]+\psi\left(\frac{\nu_j^{(k)}+p}{2}\right)-log\left(\frac{\nu_j^{(k)}+p}{2}\right) \tag{B.10}$$

which can be found, e.g., by a bisection algorithm or by quasi-Newton methods as the left hand side expression is monotonically increasing in $\nu_j^{(k)}$

B.2.2 Normal component distribution

The re-estimation formulae for the normal component distributions is straight forward, and is given by Bulla [2006] who refers to Sansom and Thomson [2001]:

$$\mu_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t) x_t}{\sum_{t=0}^{\tau-1} L_j(t)} \tag{B.11}$$

$$\sigma_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)(x_t - \mu_j)^2}{\sum_{t=0}^{\tau-1} L_j(t)} \tag{B.12}$$

Bibliography

Blattberg, R. C. and N. J. Gonedes (1974). A comparison of the stable and student distributions as statistical models for stock prices. *The Journal of Business 47*(2), 244–280.

Brockwell, P. J. and R. A. Davis (2002). Introduction to time series and forecasting. *Springer Texts in Statistics Second Edition*.

Bulla, J. (2006). Application of hidden markov models and hidden semi-markov models to financial time series. *Munich Personal RePEc Archive March*(7675), 20–25, 57–80, 99–105.

Bulla, J. (2011). Hidden markov models with t components. increased persistence and other aspects. *Quantitative Finance 11*(3), 459–475.

Bulla, J. (2013). Computational advances and applications of hidden (semi-)markov models. *Habilitation thesis*.

Bulla, J. and I. Bulla (2006). Stylized facts of financial time series and hidden semi-markov models. *Computational Statistics and Data Analysis Desember*, 2192–2209.

Bulla, J., I. Bulla, and O. Nenadić (2010). hsmm — an r package for analyzing hidden semi-markov models. *Computational Statistics and Data Analysis 54*, 611–619.

Cai, J. and J. Garrido (1999). A unified approach to the study of tail probabilities of compound distributions. *Journal of Applied Probability 36*(4), 1058–1073.

Cartella, F., J. Lemeire, L. Dimiccoli, and H. Sahli (2014). Hidden semi-markov models for predictive maintenance. *Hindawi Publishing Corporation*.

Chen, J. (2020). Conditional Value at Risk (CVaR). https://www.investopedia.com/terms/c/conditional_value_at_risk.asp.

Chung, H., E. Loken, and J. L. Schafer (2004). Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician 58*(2), 152–158.

Costa, M. and L. D. Angelis (2010). Model selection in hidden markov models: a simulation study. *Serie Ricerche* (7).

Davis, C. (2015). The skewed generalized t distribution tree package vignette.

Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business 38*(1), 34–105.

Fosler-Lussier, E. (1998). Markov models and hidden markov models: A brief tutorial. *International Computer Science Institute*.

Goodall, V. L. (2014). Statistical approaches towards analysing ungulate movement patterns in the kruger national park. *PhD-dissertation*, 161–164.

Granger, C. W. J. and Z. Ding (1995a). Some properties of absolute return: An alternative measure of risk. *Annales d'Économie et de Statistique* (40).

Granger, C. W. J. and Z. Ding (1995b). Stylized facts on the temporal and distributional properties of daily data from speculative markets. *Department of Economics, University of California, San Diego* (unpublished paper).

Greunen, J. V. (2011). Determining the impact of different forms of stationarity on financial time series analysis. *PhD-thesis*, 8–18.

Guédon, Y. (2003). Estimating hidden semi-markov chains from discrete sequences. *Journal of Computational and Graphical Statistics 12*(3), 604—639.

Jurafsky, D. and J. H. Martin (2019). Speech and language processing. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition 3rd edition draft*, 548–563.

Kent, J. T., D. E. Tyler, and Y. Vard (1994). "A curious likelihood identity for the multivariate t-distribution". *Communications in Statistics - Simulation and Computation 23*(2), 441–453.

Kenton, W. (2019). Value at Risk (VaR). https://www.investopedia.com/terms/v/var.asp.

Kon, S. J. (1984). Models of stock returns–a comparison. *The Journal of Finance 39*(1), 147–165.

Lin, X. S. (2006). Compound distributions.

Liu and Wang (2017). Decoding chinese stock market returns: Three-state hidden semi-markov model. *Pacific-Basin Finance Journal 44*.

Ma, D. (2010). Applied probability and statistics in actuarial science and financial economics - an introduction to compound distribution. https://mathmodelsblog.wordpress.com/2010/01/17/an-introduction-to-compound-distributions/.

Marin, J.-M., K. Mengersen, and C. P.Robert (2005). *Bayesian Modelling and Inference on Mixtures of Distributions*, Volume 25.

Maruotti, A., A. Punzo, and L. Bagnato (2019). Hidden markov and semi-markov models with multivariate leptokurtic-normal components for robust modeling of daily returns series. *Journal of Financial Econometrics 17*(1), 91–117.

McLachlan, G. and D. Peel (2000). Finite mixture models. *Wiley series in probability and statistics - applied probability and statistics section*, 37–50.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2018). Finite mixture models. *Annual Review of Statistics and Its Application 6*, 355–378.

McNeil, A. J., R. Frey, and P. Embrechts (2005). Quantitative risk management. *Princeton University Press 101*, 1–3, 38–44.

Murphy, K. P. (2012). Hidden semi-markov models (hsmms).

Narimatsu, H. and H. Kasai (2019). State duration and interval modeling in hidden semi-markov model for sequential data analysis. *Annals of Mathematics and Artificial Intelligence february*.

Nguyen, N. (2018). Hidden markov model for stock trading. *International Journal of Financial Studies 6*, 1–17.

O'Connell, J. and S. Højsgaard (2011). Hidden semi markov models for multiple observation sequences: The mhsmm package for r. *Journal of Statistical Software 39*(4).

O'Connell, J., F. A. Tøgersen, N. C. Friggens, P. Løvendahl, and S. Højsgaard (2011). Combining cattle activity and progesterone measurements using hidden semi-markov models. *Journal of Agricultural, Biological, and Environmental Statistics 16*(1), 1–16.

Palachy, S. (2019). Stationarity in time series analysis. https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322.

Peel, D. and G. J. MClachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*, 339—-348.

Phoong, S.-Y. and M. T. Ismail (2014). A study of finite mixture model: Bayesian approach on financial time series data. *AIP Conference Proceedings 1605*, 805–808.

Picard, F. (2007). An introduction to mixture models. *Statistics for Systems Biology Group* (7).

Pitts, S. M. (1994). Nonparametric estimation of compound distributions with applications in insurance. *Annals of the Institute of Statistical Mathematics 46*(3), 537–555.

Pohle, J., R. Langrock, F. M. van Beest, and N. M. Schmidt (2017). Selecting the number of states in hidden markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural Biological and Environmental Statistics 22*(3), 1–24.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 257–286.

Ruppert, D. and D. S. Matteson (2015). Statistics and data analysis for financial engineering. *Springer Texts in Statistics Second Edition*, 5–8, 553–559.

Rydén, T., T. Teräsvirta, and S. Åsbrink (1998). Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics 13*(3), 217–244.

Sansom, J. and P. Thomson (2001). Fitting hidden semi-markov models to breakpoint rainfall data. *Journal of Applied Probability January*(38).

Shevchenko, P. V. (2010). Calculation of aggregate loss distributions. *The Journal of Operational Risk 5*(2), 3–40.

Stamp, M. (2018). Introduction to machine learning with applications in information security. *Chapman Hall/CRC Machine Learning Pattern Recognition Series 1st Edition*, 7–28.

Suda, D. and L. Spiteri (2019). Analysis and comparison of bitcoin and s and p 500 market features using hmms and hsmms. *MDPI*.

Tolver, A. (2016). An introduction to markov chains. *Lecture notes for Stochastic Processes First printing*, 15–18.

Willmot, G. and X. Lin (2001). *Compound distributions*, Volume 156.

Yu, S.-Z. (2010). Artificial intelligence. *Elsevier 174*(2), 215–243.

Zucchini, W., I. L. MacDonald, and R. Langrock (2016). Hidden markov models for time series. an introduction using r. *Monographs on Statistics and Applied Probability 150 Second Edition*, 165–185.